Cryo-electron microscopy structure determination using the COSMIC^2 science gateway

Structural biology is in the midst of a revolution. Instrumentation and software improvements have allowed for the full realization of cryo-electron microscopy (cryo-EM) as a tool capable of determining atomic structures of protein and macromolecular samples. These advances open the door for solving new structures that were previously unattainable, which will soon make cryo-EM a ubiquitous tool for structural biology worldwide, serving both academic and commercial purposes. However, despite its power, new users to cryo-EM face significant obstacles. One major barrier consists of the handling of large datasets (10+ terabytes), where new cryo-EM users must learn how to interface with the Linux command line while also dealing with managing and submitting jobs to high performance computing resources. To address this barrier, we have developed the COSMIC^2 Science Gateway as an easy, web-based, science gateway to simplify cryo-EM data analysis using a standardized workflow to run on XSEDE's [1] supercomputers. This gateway will lower the barrier to high performance computing tools and facilitate the growth of cryo-EM to become a routine tool for structural biology. This gateway is operational and will be open for beta testing shortly, requiring that we provide users with XSEDE SUs through our science gateway as a community account. Therefore, we are requesting allocations on Comet, Comet GPU, XStream, and Oasis to provide up to 200 users access to cryo-EM software on supercomputer resources through the COSMIC^2 science gateway.

1. Proposal summary

We have built a new science gateway for cryo-EM: COSMIC^2 (*C*ryo-EM *O*pen *S*ource *M*ultiplatform *I*nfrastructure for *C*loud *C*omputing) (cosmic-cryoem.org). This science gateway will alleviate the computational burden of cryo-EM by providing users access to Comet for data analysis while also removing the command-line from data processing. This will ensure the spread of cryo-EM to decrease the time it takes to determine a cryo-EM structure. We have built the gateway using grants from XSEDE ECSS and the Science Gateways Community Institute to provide FTEs to developed at SDSC, allowing us to build the COSMIC^2 gateway in less than 12 months by using the successful CIPRES science gateway as a template [2].

We are requesting an XSEDE allocation for a community account that will be utilized by the COSMIC^2 gateway. In order to move the COSMIC^2 gateway into production mode, we are requesting an XSEDE allocation on Comet, Comet GPUs, and XStream perform. For this to occur, we will benchmark the GPU-implementation of Relion-2.0, incorporate into the gateway submission pipeline, release to alpha-testers, and, finally, the public.

Project goals:

1. **Incorporation of Relion-2.0 into COSMIC2 gateway**
   - Benchmark performance using K80 and P100 GPU nodes
2. **Utilize Comet GPU nodes for cryo-EM structure determination with COSMIC^2 gateway**
   1. Alpha testing
   2. Beta testing
   3. Public release
3. **Compare performance of Comet GPU vs. XStream for Relion-2.0 jobs**

2. Background for proposal

2.1 Cryo-EM as a fast growing field of structural biology

Structural biology, the field devoted to understanding the chemical underpinnings of life, was born in the 1950's with the determination of the atomic-resolution structures of DNA and the first two proteins—myoglobin and hemoglobin—by X-ray crystallography. Structural biology has been dominated by two techniques since its early days: X-ray crystallography (1950's), and Nuclear Magnetic Resonance (NMR) (1980's). With over 120,000 structures deposited in the Protein Data Bank (PDB), it is impossible to overstate the impact crystallography and NMR have had on our understanding of basic biological function and on human health.

These two techniques do, however, suffer from some limitations that have prevented their application to many multi-component biological structures. These limitations have become more apparent as we have realized, over the last decade or two, that cellular functions are carried out not by the isolated proteins envisioned by early molecular biology, but rather by very large, and highly coordinated assemblies of many individual components. Understanding how these assemblies work, and what their emergent properties are, has become a driving force in contemporary structural biology.

A "third" structural biology technique, cryo-EM, has recently come to the forefront of structural biology due to its ability to 'fill in' this missing gap of knowledge. Its arrival was born over decades of technological advances, culminating in atomic protein structures only in the last few years (see [3] for a historical perspective). These advances have led to an explosion in the field [4], with weekly high-resolution cryo-EM structures being published in top-tier journals. This led to the journal Nature Methods to name cryo-EM "Method of the Year" in 2015 [5].

Year-to-year growth[#] in PDB depositions from the three major structural techniques

	2013	**2014**	**2015**	**2016***
X-ray crystallography	6.9%	1.5%	-2.3%	15.1%
NMR	-5.8%	9.5%	-20.8%	0.4%
Cryo-EM	81.5%	59.3%	14.9%	40.5%

[#] Relative to previous year.
* Data for 2016 were calculated by extrapolating the current number of PDB depositions to the full 12 months (which likely underestimates the total number).

In order to take advantage of this new structural biology tool, researchers worldwide are collecting and analyzing unprecedented amounts of data. Per experiment, an individual scientist will process up to 15 terabytes of data from a single dataset collected over the course of a week. Given that most research laboratories comprise 8 – 15 scientists, typical laboratories are now responsible for managing and analyzing hundreds of terabytes of data. Furthermore, with this data in hand, these same scientists now require 50,000 core-hours per structure, necessitating the use of high performance computing (HPC) resources. This 'big data' problem has slowed the spread of the cryo-EM technique, requiring that new users must learn and master cumbersome command line tools and cluster submission routines.

2.1.1 High performance computing requirements for cryo-EM

Unlike other structural biology techniques (X-ray crystallography and NMR), cryo-EM requires the collection and analysis of large detests (> 5TB per dataset). Due to the large dataset sizes, cryo-EM requires the use of high performance computing resources (HPC), as structures typically required 50,000 - 100,000 CPU core-hours per structure.

The requirement for HPC has negatively impacted the spread of cryo-EM, as only large universities and institutions that invest in computational infrastructure are able to utilize cryo-EM for structure determination. This need has been addressed by XSEDE, providing cryo-EM data analysis software on both Gordon and Comet. My research has been positively impacted through access to these supercomputers through XSEDE allocations MCB160079, MCB160104, and MCB140257.

2.1.2 Cryo-EM utilization of GPU-accelerated software

During 2016, the software developers for Relion implemented GPU-accelerated code that now allows users to take advantage of the power of GPU processors for cryo-EM, using version 2.0 [6]. This has led to the further adoption of Relion and many research universities and institutes investing in GPU processors for cryo-EM structure determination, due to the reduced computational overhead to to set up GPU clusters instead of the equivalently sized CPU cluster.

Despite this spread, GPU architectures and setup remains a hurdle for many research laboratories, considering that these GPU-based machines require server racks, chilling systems, and IT administrators. These considerations have slowed down the spread of cryo-EM to many labs due to inability to set up a system for users.

2.2 COSMIC^2 science gateway

To address the HPC demands of cryo-EM while also facilitating its spread, we have built the COSMIC^2 'science gateway'. As a science gateway, COSMIC^2 will provide users access to HPC resources without the need of learning linux and command line functions. This will be critical to the spread of cryo-EM, as almost all new users for cryo-EM are biochemists who have never used a command line before, let alone HPC cluster submission environments. COSMIC^2 is hosted at SDSC and has already successfully submitted test cryo-EM jobs from the web interface to Comet at SDSC, demonstrating that our science gateway is nearing completion.

The COSMIC^2 gateway was built using the CIPRES science gateway framework [2], which allowed us to quickly start customization of the processing pipeline, building a pilot version within 12 months. The work was funded by an XSEDE ECSS grant (XSEDE MCB140257) in addition to a grant from the Science Gateways Community Institute (NSF Grant ID ACI-1547611), which allowed for 25% FTE work from developers at SDSC. Below we describe the implementation of the gateway.

2.2.1 COSMIC^2 architecture

Briefly, the COSMIC^2 gateway is hosted on a virtual machine at SDSC, allowing our gateway to serve a web server while also interfacing with the underlying Comet OASIS filesystem and job

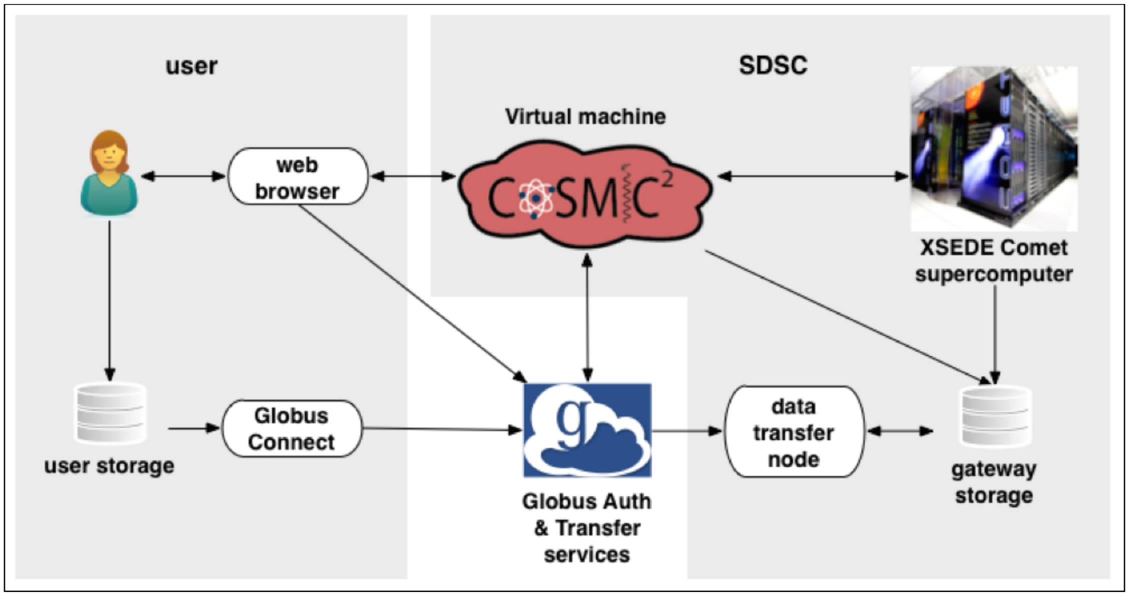


Figure 1 - Architecture of the COSMIC² science gateway. As a virtual machine hosted on SDSC, the gateway will display a web site that will direct users for authentication and data transfer to the project storage on Comet. After upload, users will be able to manage projects and data files on the gateway.

submission environment. By using Globus services for user authentication and data movement, users will be able to easily transfer large datasets to the COSMIC² project storage directory, located on Comet. (Figure 1). See below for more information regarding Globus.

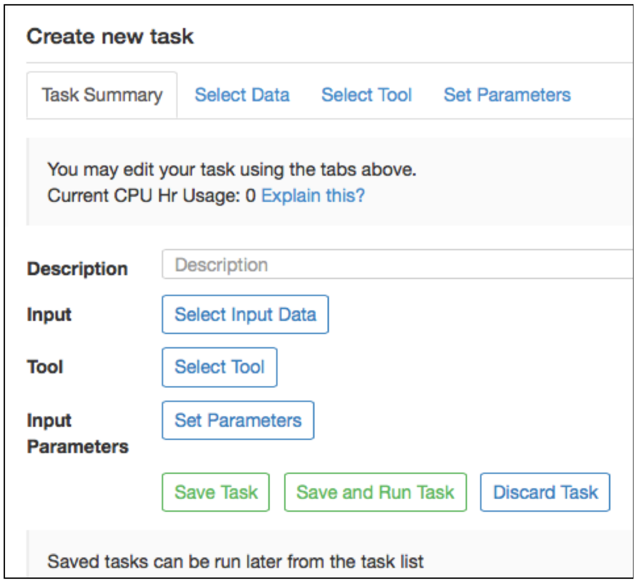
2.2.2 Using Globus services for moving large datasets

Due to the terabyte-sized data the COSMIC² gateway will need to handle from a user's local lab storage, we are incorporating Globus services developed by the University of Chicago's Computation Institute. Specifically, we will take advantage of the Globus Transfer service which provides fire-and-forget, high performance file transfers with automatic fault recovery. The transfer service utilizes Globus Auth for user identity management that can broker authentication between resources and accounts. Globus Auth builds upon the OAuth2 and OpenID Connect specifications to enable standards-compliant integration and supports identity federation models that enable diverse identities to be linked together (e.g. XSEDE, universities, Google), while also providing short-term delegated access tokens which our gateway can use to access other services on the user's behalf if needed.

An important benefit of integrating Globus services into the gateway is the ability to leverage user authentication through Globus Auth. In order to use Globus Auth, our gateway uses HTTPS in combination with a custom login utilizing Globus Auth's REST-style API. When users want to login to COSMIC², they will be redirected to the standard Globus login page to select their identity provider, authenticate against this identity, and give consent for the gateway to access their user profile information (e.g. full name, institution, email address). Upon successful authentication, the user will be automatically redirected back to the gateway where a new account will be created, if one does not already exist, and the user token object will be stored in the gateway database.

2.2.3 COSMIC^2 layout

The COSMIC^2 web server displays user projects and data based on the CIPRES gateway data management strategy. It is centered on projects and tasks, where a given project can have multiple datasets and tasks. An example screenshot for a Task is shown in Figure 2:



Create new task

You may edit your task using the tabs above.
Current CPU Hr Usage: 0 Explain this?

Description: Description

Input: Select Input Data

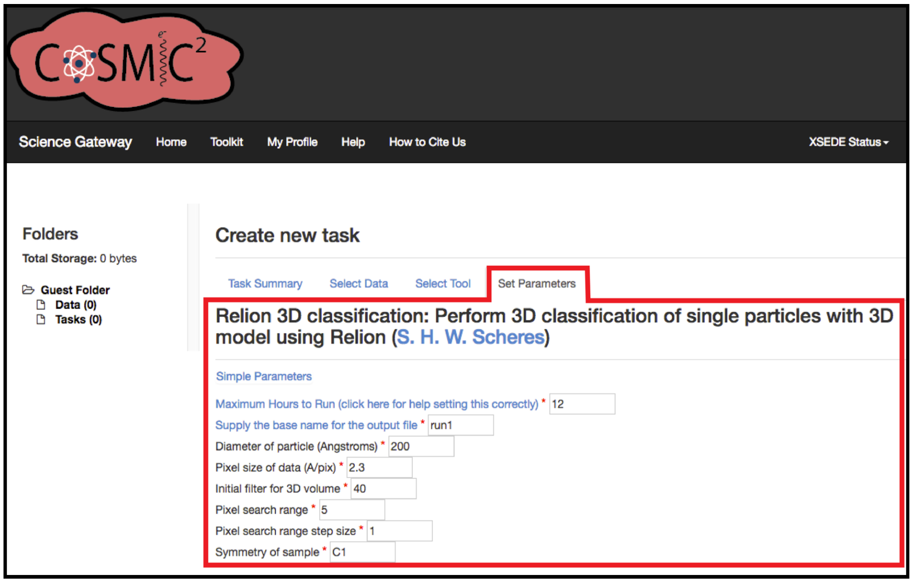
Tool: Select Tool

Input Parameters: Set Parameters

Saved tasks can be run later from the task list

Figure 2 - Example screenshot of the Task window on the COSMIC^2 website.

From this Task page, users will then select their data and the tool for analysis. Once selected, users will provide information for their analysis, with default values provided (Figure 3).



Folders
Total Storage: 0 bytes

Guest Folder
Data (0)
Tasks (0)

Create new task

Relion 3D classification: Perform 3D classification of single particles with 3D model using Relion (S. H. W. Scheres)

Simple Parameters

Maximum Hours to Run (click here for help setting this correctly) * 12
Supply the base name for the output file * run1
Diameter of particle (Angstroms) * 200
Pixel size of data (A/pix) * 2.3
Initial filter for 3D volume * 40
Pixel search range * 5
Pixel search range step size * 1
Symmetry of sample * C1

Figure 3 - Example screenshot of the parameters window for a given Task. Shown is the 3D classification page for Relion.

Once inputs are provided, the COSMIC^2 science gateway will automatically format the job into a cluster submission script with the appropriate information for nodes and processes per node. The output results will be displayed in the web server, where users can retrieve their data using the same Globus interface.

3. Project goals and timeline for COSMIC^2 science gateway

Below is a timeline for the final stages of the COSMIC2 gateway construction and the associated steps that will utilize XSEDE allocations.

April - July 2017

- Full integration of Globus into gateway to handle large file uploads
- This work will be performed by developers at SDSC through a grant from the Science Gateways Community Institute
- Incorporation of Relion-2.0 GPU into gateway
- Benchmarking of Relion-2.0 GPU on Comet GPU and comparison to Comet performance.
- **SUs: 300 on Comet GPU and 50,000 on Comet**

August - October 2017

- Alpha testing of gateway to early adaptors. 11 users have already signed up to help troubleshoot and provide feedback
- We expect users to perform 5 - 10 analysis runs, where each run likely requires 10 SUs. Therefore: 11 users x 10 runs x 10 SUs = 1,100 SUs
- **SUs: 1,100 on Comet GPU**

November 2017 - March 2018

- Beta testing will open for a large pool of users (up to 50)
- We expect each user to perform 10 analysis runs: 10 jobs x 50 users x 10 SUs/job = 5,000 SUs
- We will also test Relion-2.0 performance on XStream GPU cluster to help to decide whether we should also submit COSMIC^2 jobs to Xstream. The XStream allocation will be used to perform 10 runs from 10 separate projects, each requiring 10 SUs.
- **SUs: 5,000 on Comet GPU ; 1,000 on XStream**

April - June 2018:

- Open science gateway to the public. We expect up to 200 users that could use this gateway, although, if successful, there would be many more. 200 users x 10 jobs x 10 SUs/job = 20,000 SUs.
- **SUs: 20,000 on Comet GPU**

Total Comet: 50,000 SUs

Total Comet GPU: 26,400 SUs

Total XStream: 1,000 SUs

3.1 Incorporation of Relion-2.0 into COSMIC2 gateway

The timeline above lays out the milestones that we will cover with the COSMIC^2 gateway from June 2017 - June 2018. As described above, we will incorporate the GPU-compatible version of Relion (version 2.0) that will replace version 1.3, which was previously CPU-only. During this allocation, we will update the gateway submission program to use Relion-2.0 on Comet GPU nodes.

3.1.1 Benchmark performance using K80 and P100 GPU nodes

Since Code Performance and Scaling has not been performed with Relion-2.0 GPU on Comet GPU nodes, we will submit standardized cryo-EM analysis jobs that will benchmark the performance on both K80 and P100 GPUs. This will be critical for the gateway to automatically determine the number of GPUs per job, and also contribute information to the Comet GPU computing community.

4. Justification

As a science gateway, COSMIC^2 will remove barriers that slow down the time it takes users to solve cryo-EM structures. In doing so, it will be the first centralized data processing server for cryo-EM, allowing for users of all backgrounds to begin determining structures that have impacts in all areas of biology. Despite having a functioning gateway, we will need a stable pool of SUs that we will use through this gateway as a part of a community account for distribution to gateway users.

In order to deliver this service to the cryo-EM community, we are applying for allocations on Comet GPU (SDSC Comet GPU Nodes), Comet (SDSC Dell Cluster with Intel Haswell Processors), and XStream. Below we describe in more detail the justification for our allocation request for each XSEDE resource.

4.1 Comet GPU

Comet GPU nodes will become a central part of the analysis routines executed on the COSMIC^2 gateway. We are excited to use the new P100 NVIDIA GPUs, in addition to the K80s, to help users determine cryo-EM structures. We believe that access to GPU nodes will help to alleviate the high SU requirement that we previously had to request for CPU-based cryo-EM structures (~50,000 SUs per structure). By only requiring 5 - 10 SUs per structure, we believe it will be more efficient for us to manage these GPU nodes than large amount of Comet CPUs.

4.2 Comet

We are requesting a small allocation on Comet (Intel Haswell CPUs) to serve as a comparison to our work on Comet GPU nodes. Since the Relion-2.0 code can be run with or without GPU acceleration, we will compare the performance of the Comet GPU with the CPU performance on

Comet. This comparison, which will include typical wait times, will be critical for deciding the direction to move for future allocations.

4.3 XStream

Finally, we have also requested SUs on XStream to test the functionality of Relion-2.0 GPU code and how it compares Comet GPU nodes. Specifically, we will look at how queueing times and performance compares between the two supercomputers. This comparison will be important as we could run jobs remotely on XStream from the Comet-hosted COSMIC^2 science gateway if we need further computing power.

Other computational resources

In addition to the Comet and Comet GPU allocations, we are requesting 50,000 GB of storage on data oasis to be used for data staging during processing. Given that typical datasets will be 100 - 1000 GB per user, we believe that 50,000 GB is an adequate amount of storage to begin for this project. For 50,000 GB, we believe we would manage 20 - 50 users at a time through our gateway.

References cited

1. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, Roskies R, Scott JR, and Wilkins-Diehr N. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering.* 16, 5 (Sept. - Oct. 2014), 62–74.
2. Miller, M. A., Pfeiffer, W., and Schwartz, T. “Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees” (2010): 1–8.
3. Nogales, E. and Scheres, S. H. W. “Cryo-EM: a Unique Tool for the Visualization of Macromolecular Complexity.” *Molecular cell* 58, no. 4 (2015): 677–689.
4. Kühlbrandt, W. “Biochemistry. the Resolution Revolution.” *Science* (New York, N.Y.) 343, no. 6178 (2014): 1443–1444.
5. “Method of the Year 2015” *Nature methods* 13, no. 1 (2016): 1–1.
6. Kimanius D, Forsberg BO, Scheres SHW, Lindahl E. “Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2.” *eLIFE* 2016;5:e18722.