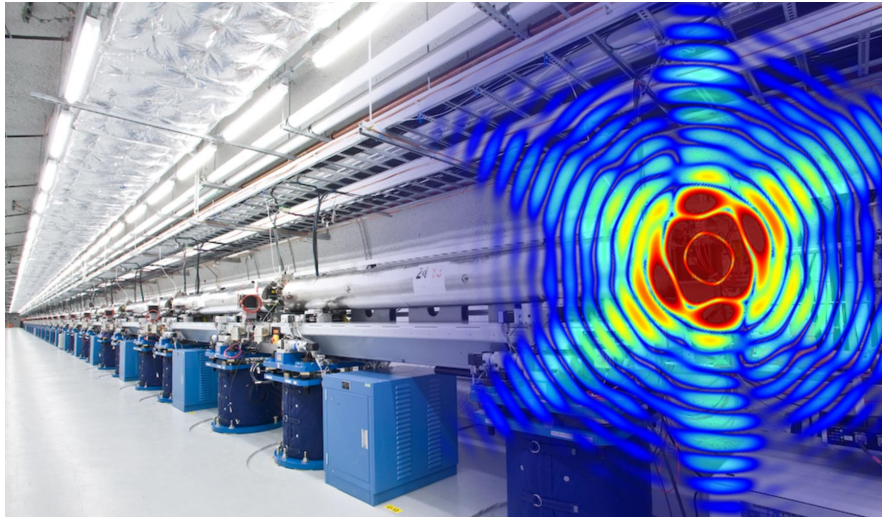


**COURSE PROJECT:  
PARALLEL GPU-BASED MARKOV CHAIN MONTE CARLO  
APPROACHES FOR PHASE RETRIEVAL WITH NOISE MODELS AND  
MISSING DATA**

CARL NETTELBLAD



BACKGROUND

This project has its origin in the field of structural biology, where the diffraction patterns arising from X-ray illumination of macroscopic crystals of biological compounds have been used as a method to determine their structure (3D coordinates for each atom) for decades. Newly developed and constructed X-ray free-electron lasers, like the 1 kilometer long Linac Coherent Light Source (LCLS) facility at Stanford University, can provide extremely powerful and extremely short bursts of X-rays. In the duration of a single pulse, the beam focused to a micron-sized spot has the same power density as all the sunlight hitting the Earth, focused to a millimeter square. With an ultra-short and extremely bright coherent X-ray pulse, a single diffraction pattern may be recorded from a *single* protein molecule, a virus particle, or a cell before the sample explodes and turns into a plasma. A femtosecond X-ray pulse can outrun slower damage processes in the sample (there is simply no time for the particle to collapse). Hence, the diffraction pattern comes from a practically undamaged object.

DESCRIPTION

Sample particles are injected individually into the X-ray beam and are intercepted randomly, in unknown orientations, by the femtosecond X-ray pulses. The data acquired by each shot represent one specific 2D projection of the sample particle. Due to the experimental geometry and physical limitations in detectors, the most central region of the pattern, as well as some other gaps, will not be recorded. Furthermore, only the signal *intensity* is recorded.

However, much of the information in a diffraction pattern is hidden in the relative *phases* of the electromagnetic wave.

#### FURTHER DETAILS

We have managed to image single virus particles using X-ray free electron lasers. The resulting images after phase retrieval have been published <sup>1</sup>.

However, as discussed in that publication, there is some uncertainty in how well those parts of the diffraction pattern that are missing are modelled, as well as how well phases are actually retrieved.

Current techniques in general use are iterative in nature, based on a tuned fix point iteration that strikes a balance between exploring the search space and tending towards local optima. The population of end results from multiple such iteration processes with different random starting points is used to assess the quality of the reconstruction. Current techniques also inherently assume that all recorded pixel values are representing the true intensity in those points, while we know that noise is involved and can even sketch what kind of non-uniform noise is expected even under ideal experimental conditions.

This approach has similarities to Markov Chain Monte Carlo techniques. We want to investigate how the phase retrieval problem would fare with a “pure” MCMC approach, based on the Metropolis-Hastings algorithm <sup>2 3</sup>. The work will be concluded on small simulated (“real space”) objects, at most 15x15 pixels. However, the recorded diffraction patterns (“Fourier space”) are 256x256 or 512x512 slices. Using the abilities in GPUs to rapidly compute the Fourier transform, as well as computing the overall probability of a specific realization, given our noise model, this should at least approach being realistic.

We will arrange access to our cluster of our 30-odd node cluster with 4 GPUs in each node (soon to be expanded).

#### AIMS

Several outcomes of specific relevance are possible from this project:

- An efficient GPU implementation for evaluation the probability of a set of true intensities given a noise model and the recorded data. This problem reoccurs in many settings in our research, beyond the MCMC approaches.
- A rudimentary Metropolis-Hastings implementation for the phasing problem.
- Evaluation of the phasing performance for this implementation on small simulated patterns, in terms of results as well as in terms of computational efficiency.
- Reproduction of phasing of experimental data on objects of such sizes with comparison against existing methods regarding the overall solution and error estimates.
- Improved noise models relevant for such experimental data.

This course project would be considered a clear success if the first two outcomes are achieved. A motivated student can choose to explore the problem further, as outlined above. Results regarding actual phasing performance in practice are far from guaranteed, but the first

---

<sup>1</sup>M. M. Seibert *et al.* Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature* **470**(7332), 2011.

<sup>2</sup>N. Metropolis *et al.* Equation of state calculations by fast computing machines. *The journal of chemical physics* **21.6**:1087-1092 (1953)

<sup>3</sup>W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications *Biometrika* **57**(1):97-109

outcome still has general applicability to our research. So has proper performance estimates for that model.

#### CONTACT

We are looking for motivated and interested students to take on this challenge. It is beneficial if you have scripting experience using Python (including numpy/scipy) and possibly MATLAB. Experience in C/C++ programming using CUDA is highly beneficial. The actual implementation will be done using a combination of CUDA, cuBLAS, cuFFT, and Thrust. Knowledge in one or several of the fields of structural biology, diffraction physics, Markov methods, and Bayesian statistics, is of course good, but in no way critical.

We are continuously looking for new ways to attack the computational challenges in this nascent area. Further possibilities for e.g. master thesis and PhD student projects might exist, independently of whether this specific approach is successful or not.

(C. Nettelblad) LABORATORY OF MOLECULAR BIOPHYSICS, DEPARTMENT OF CELL AND MOLECULAR BIOLOGY, UPPSALA UNIVERSITY, SWEDEN.

*E-mail address:* [carl.nettelblad@icm.uu.se](mailto:carl.nettelblad@icm.uu.se)