# Probabilistic Graphical Models and Machine Learning for 3D Image Analysis Across Domains

[1]Talita Perciano, [1]Wes Bethel, [1]David Camp, [2]Hank Childs, [3]Karen Davies, [1]Colleen Heinemann, [1]Hari Krishnan, [4]Carolyn Larabell, [2]Brent Lessley, [5]Dula Parkinson, [3]Nicholas Sauter, [6]James Sethian, [1]Daniela Ushizima, [3]Peter Walian, [1]Chao Yang, [1]Li Zhou

[1]Computational Research Division, LBNL; [2]University of Oregon; [3]Molecular Biophysics & Integrated Bioimaging Division, LBNL; [4]Department of Anatomy, UCSF; [5]Advanced Light Source, LBNL; [6]University of California Berkeley

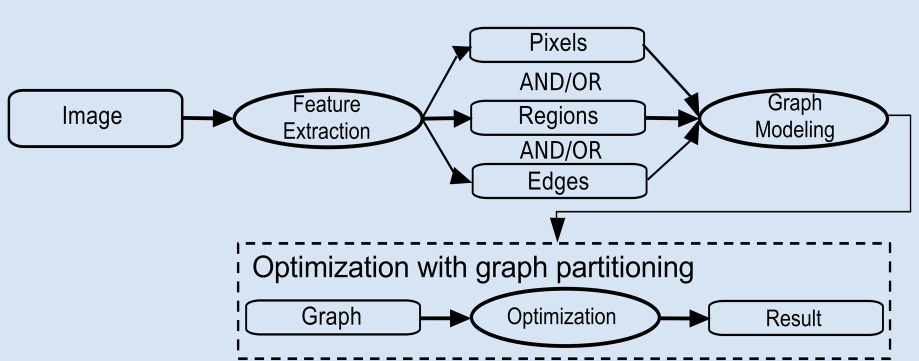
BERKELEY LAB

U.S. DEPARTMENT OF ENERGY Office of Science

- DOE research across science domains is increasingly reliant on **image-based data from experiments**
- Need for **new analysis tools** that help scientists **uncover relevant, but hidden, information in digital images**
- Better exploit the scientific value of a broad array of **high resolution, multidimensional and multiscale datasets**
- We are developing state-of-the-art research using **probabilistic graphical models, pattern recognition and machine learning** techniques, applied to a **diverse range of science problems**
- Target the **portability and efficiency** of our computational approaches, which are being tested using computational resources at NERSC

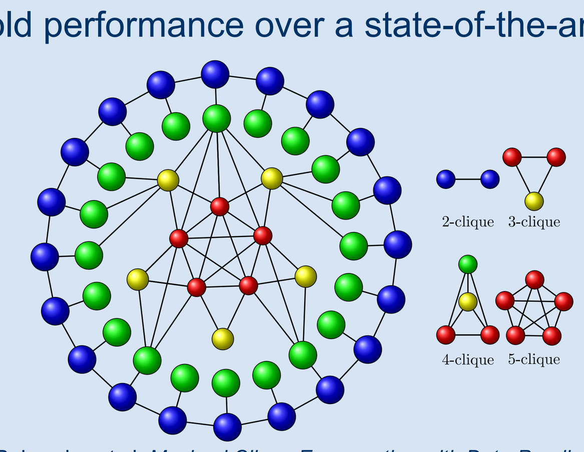
## Parallel Markov Random Fields - PMRF

- Probabilistic Graphical Model with graph partitioning
- Optimization function including physical constraints
- Based on maximal cliques of the graph



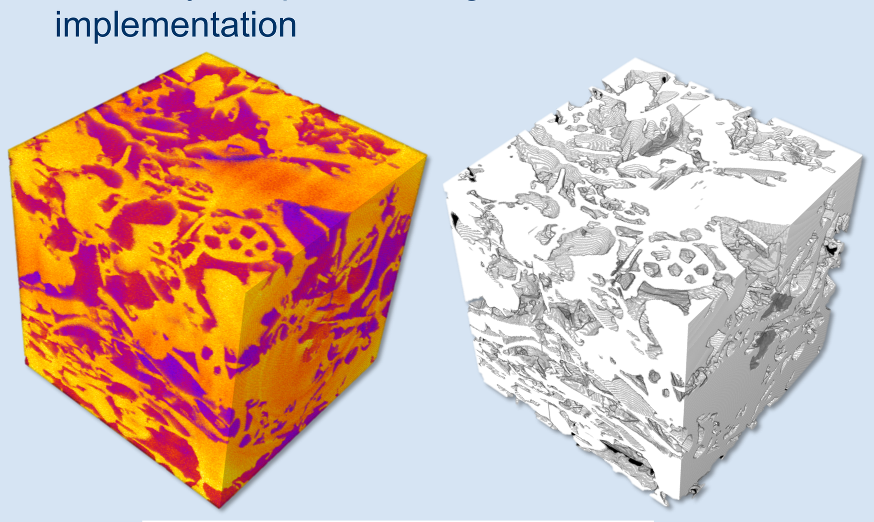

T. Perciano et al. *Reduced-complexity image segmentation under parallel Markov random field formulation using graph partitioning*, ICIP 2016.

- New maximal clique enumeration (MCE) algorithm using data parallel primitives
- A 10-fold performance over a state-of-the-art approach




B. Lessley et al. *Maximal Clique Enumeration with Data-Parallel Primitives*. IEEE LDAV 2017, Oct. 2017.

- Distributed memory version of PMRF using MPI
- High segmentation accuracy (>98%) and higher efficiency compared to original PMRF implementation



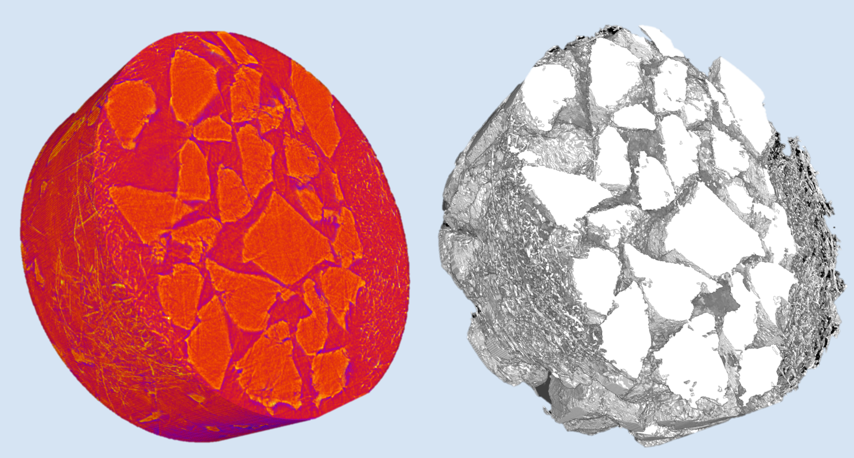
| Number Processes | Synthetic Time (in seconds) | Experimental Time (in seconds) |
|---|---|---|
| **1** | 70.6 | 633.9 |
| **2** | 37.4 | 340.4 |
| **4** | 20.2 | 129.9 |
| **12** | 8.3 | 50.3 |
| **24** | 4.7 | 38.4 |
| **48** | 2.7 | 33.5 |
| **96** | 1.7 | 19.1 |
| **192** | 1.4 | 11.9 |
| **384** | 1.4 | 7.6 |
| **768** | 1.3 | 6.9 |

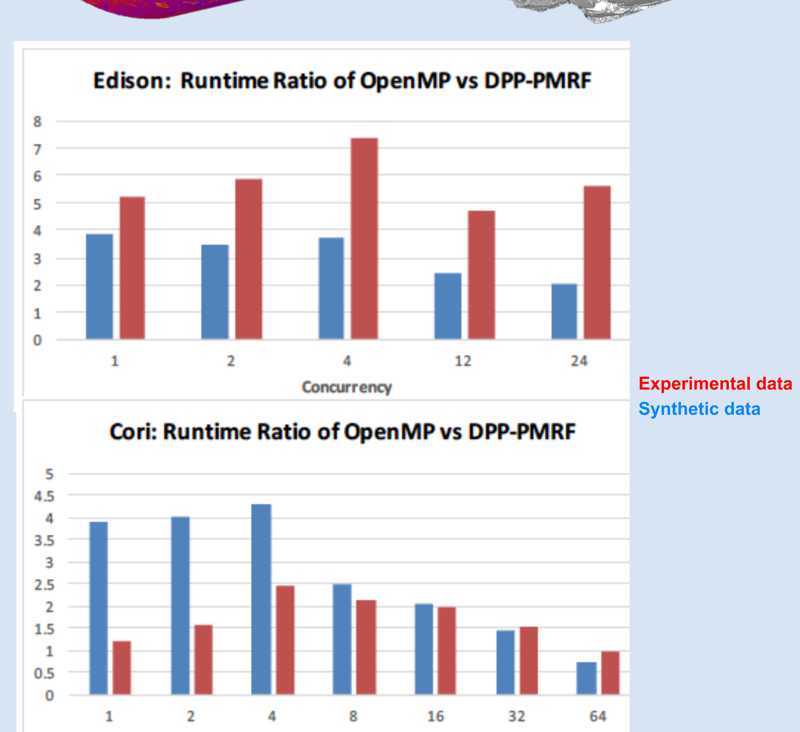
$$E(n,p) = \frac{C^*(n)}{C(n,p)}$$

$$R(n,p) = \frac{n}{T(n,p)}$$

C. Heinemann el al. *Distributed Memory Parallel Markov Random Fields Using Graph Partitioning*. IEEE BigData/BigGraphs 2017

- PMRF reformulation using data parallel primitives
- An order-of-magnitude performance increase on CPU and GPU platforms

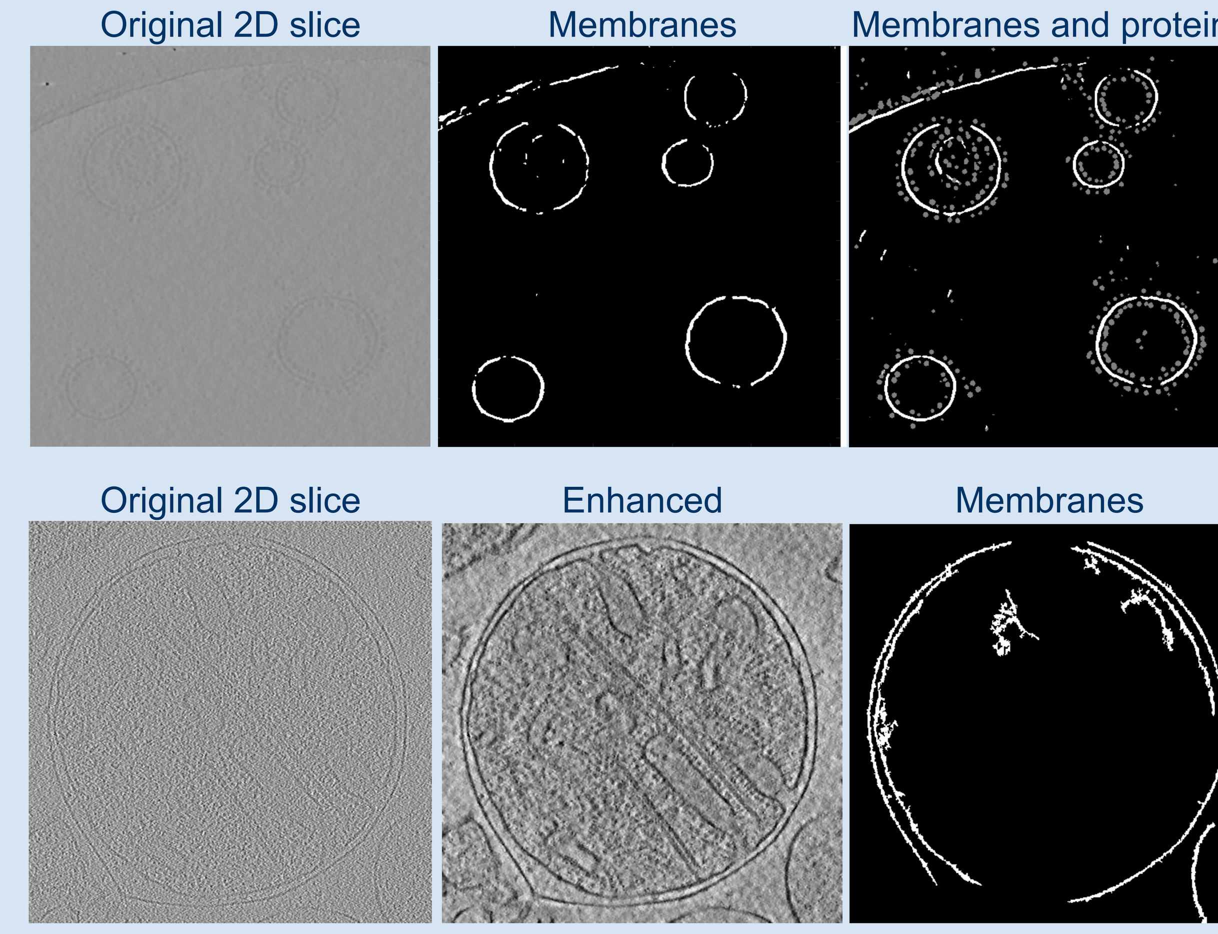





B. Lessley et al. *DPP-PMRF: Rethinking Optimization for a Probabilistic Graphical Model Using Data-Parallel Primitives.* IEEE LDAV 2018, Oct 2018.

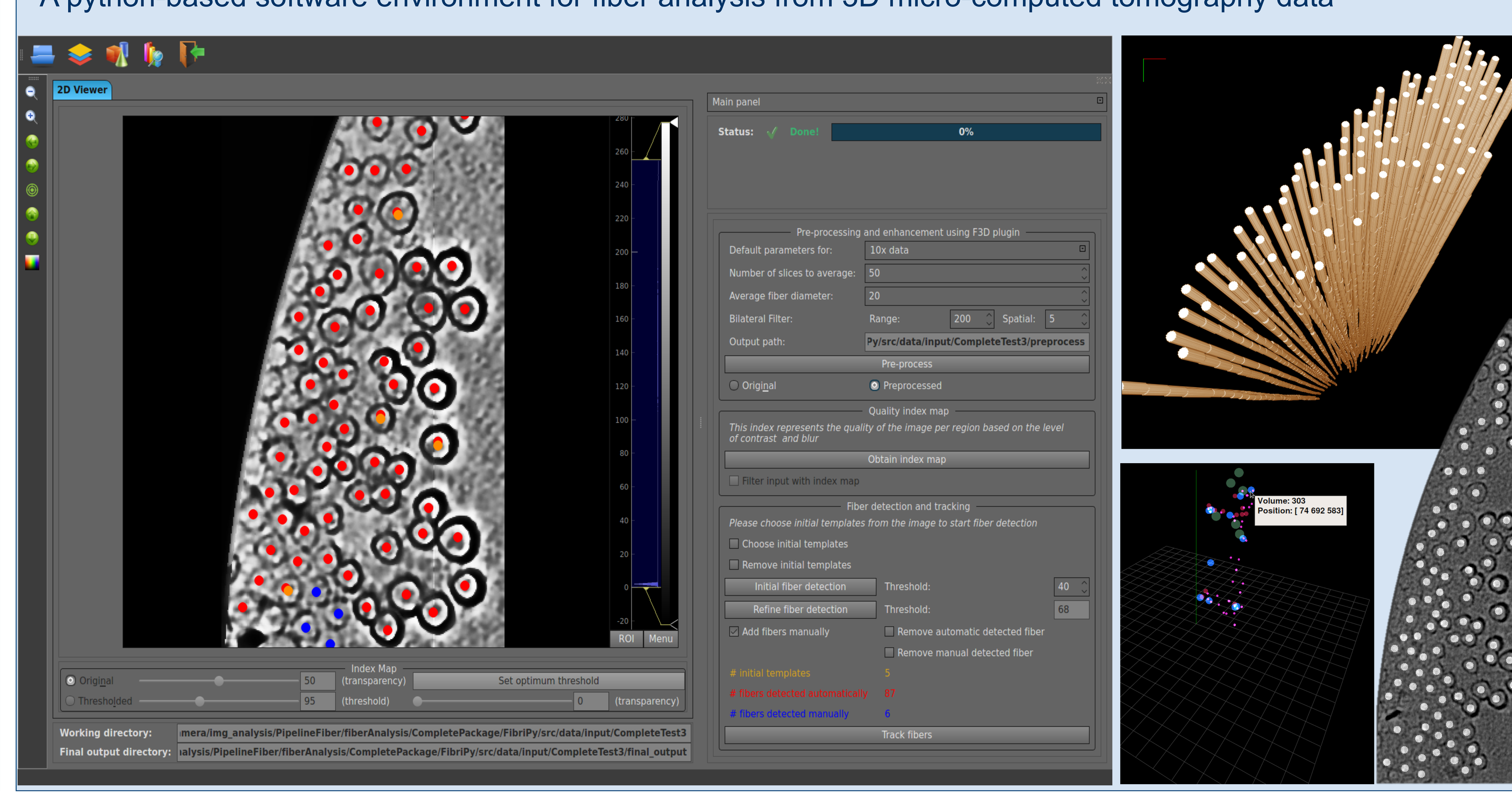
## Machine Learning for Tomogram Segmentation - MLTS

- **Problem**: segmentation of cell tomograms into organelles
- **Data**: cryo-electron tomography, soft and hard x-ray tomography
- **Contributions**: new synthetic training data, state-of-the-art enhancement algorithm, unsupervised and supervised segmentation approaches using deep learning and graph-based techniques
- **Uniqueness**: incorporate domain-specific knowledge into the machine learning techniques, along with user-generated guidance for active learning




## FibriPy

A python-based software environment for fiber analysis from 3D micro-computed tomography data



## Research Details

**PMRF**

- **Unique breadth-first parallel MCE algorithm** using DPPs
- First algorithm to perform **parallel MRF parameter estimation and optimization**
- MPI, DPP and OpenMP implementations

**MLTS**

- State-of-the-art enhancement algorithm
- Unique synthetic training data
- **Deep learning and graph-based approaches** for tomogram segmentation using **prior information** and **active learning**

**FibriPy**

- Unique **image enhancement** and **feature tracking** optimized for high-resolution images
- **Deep learning module** enables high accuracy detection and creation of training data

## Scientific Achievement

**PMRF**

- New method for graph partitioning **scales image segmentation** to run on large DOE HPC platforms, and with high accuracy
- **An order-of-magnitude performance increase** on CPU and GPU platforms results from algorithmic reformulation using data-parallel primitives

**MLTS**

- **Combined unsupervised and supervised** approach for tomogram segmentation using **graph modeling**, **deep learning** and **active learning**

**FibriPy**

- Software framework that enables **accurate fiber detection** and **measurement** from 3D imagery

## Significance and Impact

**PMRF**

- **Accelerating image analysis** time-to-solution for DOE science projects challenged by an ever-growing *data tsunami*
- **Portability** and **performance** on rapidly evolving computer architectures, into the exascale regime and beyond into Post-Moore's Law
- Applications: from computational chemistry and bioinformatics, to cybersecurity

**MLTS**

- Enable **a high-throughput**, **deep-mining** approach to biological imaging
- Analysis to quantitative conclusions

**FibriPy**

- Development of advanced materials through **automatic image analysis**
- Improve quality control of materials in the **manufacturing processes**

## Acknowledgements

- Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. DOE under Contract No. DEAC02-05CH11231
- "Towards Exascale: High Performance Visualization and Analytics," program manager Dr. Lucy Nowell
- "Scalable Data-Computing Convergence and Scientific Knowledge Discovery", program manager Dr. Laura Biven
- Center for Applied Mathematics for Energy Related Applications (CAMERA)
- U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Award Number 14-017566, Program Manager Lucy Nowell
- This research used resources of the NERSC Center
- DOE LDRD program project number 106249 entitled "Interactive Machine Learning for Tomogram Segmentation and Annotation"

CAMERA Applied Math · CRD · UCSF University of California San Francisco · ALS · BioSciences · UNIVERSITY OF OREGON · BIDS BERKELEY INSTITUTE FOR DATA SCIENCE