BERKELEY LAB

Bringing Science Solutions to the World

BIOLOGICAL SCIENCES SOFTWARE

Tools to accelerate discovery and products-to-market

INTRODUCTION

This brochure presents a sample of Berkeley Lab software tools that can accelerate discovery and hasten products-to-market in the biosciences. Some of this software is available open source and some through commercial licenses.

The primary contributors to these software tools include Berkeley Lab's Biosciences Area divisions, as well as the Lab's Computational Research Divisions and the National Energy Research Scientific Computing Center (NERSC). Other contributors include Sandia National Laboratories and Lawrence Livermore National Laboratory. Development of these software packages was only possible due to significant investments by the U.S. Department of Energy in world class facilities and research projects and programs, such as the Advanced Light Source, the Joint Genome Institute, the Joint BioEnergy Institute, and the Agile BioFoundry.

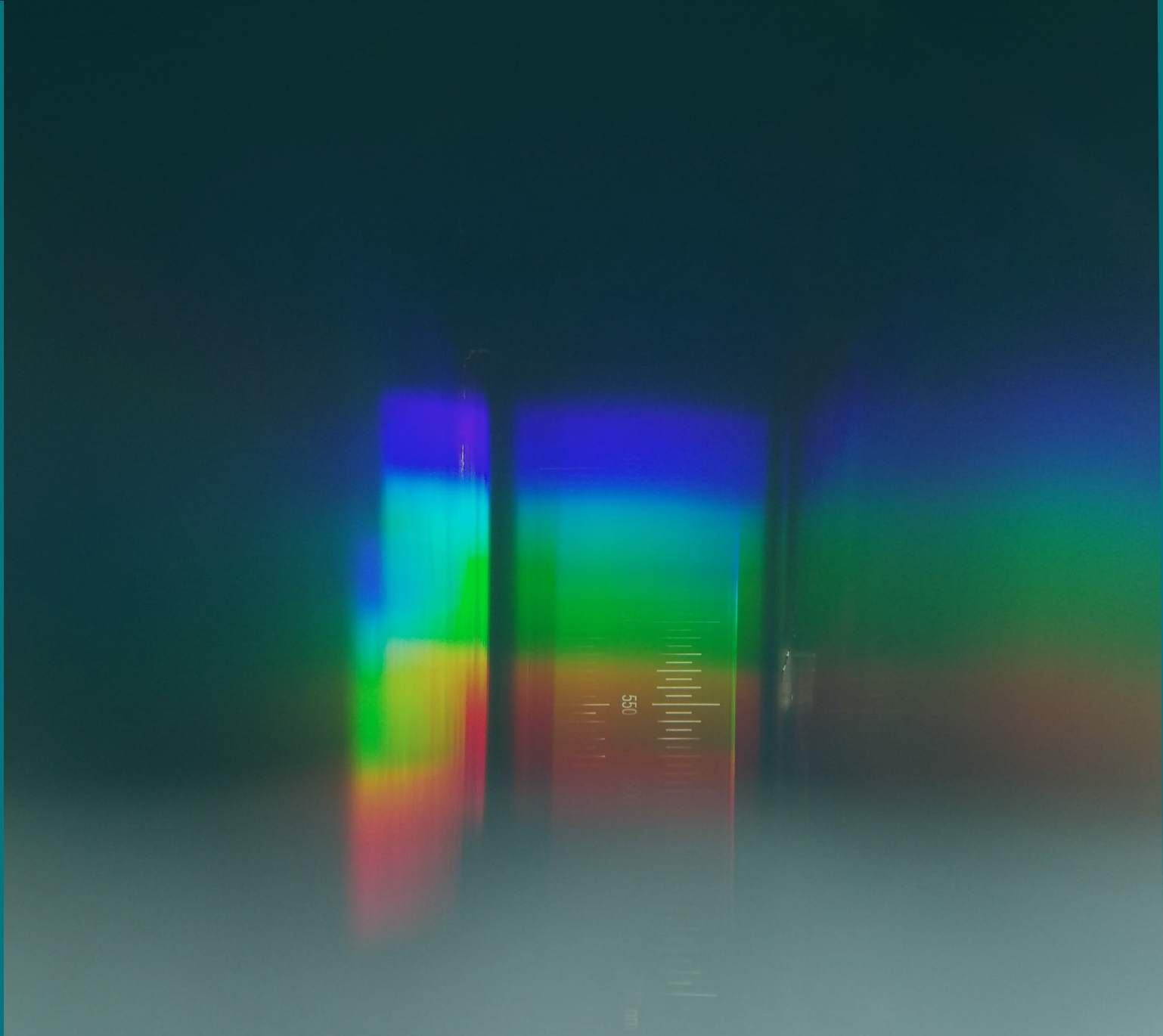
Contact: ipo@lbl.gov
Website: ipo.lbl.gov
LinkedIn: Berkeley Lab IPO

Contributing Berkeley Lab Areas, Divisions, and Partners

TABLE OF CONTENTS

Go directly to pages by clicking on titles

***open source**



IMAGING

BASTet and OpenMSI Arrayed Analysis Tools for Mass Spectrometry

2016-107

Oliver Ruebel, Ben Bowen, Tristan de Rond, Markus de Raad

APPLICATIONS OF TECHNOLOGY

- Analysis and storage for the OpenMSI project

BENEFITS

- An easier, streamlined computational method for handling mass spectrometry (MS) data

BACKGROUND

Mass spectrometry imaging (MSI) enables high-resolution spatial mapping of biomolecules in samples and the analysis of tissues from plants and animals, microbial interactions, high-throughput screening, drug metabolism, and a host of other applications. Although it is an increasingly important tool in life science research and development, its use has been limited for lack of software tools to visualize, manipulate, store, and transfer MS data.

TECHNOLOGY OVERVIEW

Researchers at LBNL have developed an advanced software library known as Berkeley Analysis and Storage Toolkit (BASTet) that serves as the analysis and storage library for the OpenMSI project. Written in Python, BASTet is an integrated framework for:

- Storage of spectral imaging data and derived analysis data
- Provenance of analyses
- Integration and execution of analyses via complex workflows
- Defining interfaces to enable developers to directly integrate their analysis with OpenMSI's web-based viewing infrastructure without having to know OpenMSI

Additionally, researchers have addressed the limited use of MS with the development of the OpenMSI Arrayed Analysis Tool (OMAAT). This novel computational software method addresses the challenges of analyzing spatially defined samples in large MSI datasets, by providing support for automatic sample position optimization and ion selection. OMAAT's algorithm automatically finds outlier data. It is written in Python with an accompanying Jupyter (formerly iPython) notebook and is fully integrated with OpenMSI.

SPECT: Software for Dynamic Imaging

3010

Rostylav Buchko, Grant T Gullberg, Bryan W Reutter, Yuval Zelnik

APPLICATIONS OF TECHNOLOGY

- Dynamic diagnostic imaging
- Research on tissue perfusion and metabolism

BENEFITS

- Accurate quantitation of coronary flow reserve, tissue perfusion, and viability
- Applicable with slowly rotating SPECT scanners or with PET scanners

TECHNOLOGY OVERVIEW

Scientists at Berkeley Lab have developed a computer program that uses data from single-photon emission computed tomography (SPECT) to quantify and create images of dynamic blood flow in living tissue. This tool will improve noninvasive diagnosis and assessment of diseases using SPECT scanners, which are more common and less expensive than positron emission tomography (PET) scanners. The software generates system matrices that use SPECT data to provide accurate arterial input functions—mathematical functions describing the flow of blood into tissue over time. These functions can be used to quantify regional tissue perfusion and viability, as well as coronary flow reserve. The software can also be applied directly to PET data for cases when these scans are available.

The scientists tested the system by applying it to data from three patients who underwent PET scans of the heart. These data were used to simulate projection data that would be obtained from the same patients with a slowly rotating double-headed SPECT scanner. The simulated SPECT data was then processed with both existing software, which uses a standard three-dimensional (3D) technique at multiple time points, and the new software, which uses a spatiotemporal (4D) technique. The results from each type of software were compared to the original PET results as an indicator of accuracy. The 4D technique was more accurate than the 3D technique in measuring arterial input functions. The 4D program also produced accurate time-activity curves for blood and myocardium and used these to estimate the tissue uptake rate parameter, K1. In addition, the software was able to compensate for errors introduced by radiotracer signal attenuation, collimation-induced blurring, and changes in radiotracer distribution during camera rotation.

Dynamic imaging of tissue to visualize and quantify blood flow and regional metabolic rates is useful in assessing cardiac and other diseases, as well as normal physiology. While this type of imaging can be performed with PET scanners, these scanners are rare, expensive, require a nearby cyclotron or generator, and can only image one type of isotope at a time. MRI images may also be useful but not all patients can undergo MRI scans. SPECT scanners are more common, convenient, and less expensive than PET scanners and more readily tolerated than MRI scans. However, SPECT scanners have limited capability for dynamic imaging, because, unlike PET scanners, their cameras must rotate slowly about the body as the radiotracer distribution changes. The Berkeley Lab software overcomes these limitations, allowing the more accessible SPECT scanners to provide objective, quantitative data for diagnosis and research.

LABELIT : Software for Macromolecular Diffraction Data Processing

1960

Nicholas K. Sauter

APPLICATIONS OF TECHNOLOGY

- Output from Berkeley Lab's LABELIT software program for rapid processing of X-ray diffraction images.
- Automatically index X-ray diffraction images
- Quickly summarize the diffraction quality
- Determine Laue symmetry from a partial dataset

BENEFITS

- Makes visual inspection of images unnecessary
- Better algorithms for spot picking
- More robust indexing procedures
- Suitable for automatic scripting without requiring graphical interaction

TECHNOLOGY OVERVIEW

The Lawrence Berkeley Laboratory Indexing Toolbox (LABELIT) expedites the processing of X-ray diffraction images at the beamline, in situations such as raster screening where the pace of data collection outstrips the ability to analyze each new sample with conventional graphics-based tools.This software is available for commercial end-user licensing.

LABELIT is the diffraction data indexing program of choice for automating a production line. Visual inspection of the images is no longer necessary–LABELIT builds upon the experience gained from thousands of contributed datasets, to give the correct lattice even under unusual conditions. Matching to a known symmetry and unit cell is possible.

There are three main functions of LABELIT: 1) analysis of the Bragg spots to produce an initial estimate of the resolution limit, along with a summary of the diffraction artifacts such as ice-rings; 2) indexing, and a preliminary identification of the lattice symmetry; and 3) integration of a partial dataset to obtain the true Laue symmetry, which is necessary for planning the remaining data acquisition. Additional functions allow the indexing results to be passed to the program MOSFLM for data integration. The program is easily controlled by command-line or Python-language interfaces, allowing the Bragg spot analysis and autoindexing results to be incorporated into one's own scripts.

PHENIX: Software for Determining Macromolecular Structure Using Crystallographic or Cryo-EM Data

IB-1770

Lawrence Berkeley National Laboratory (Paul Adams group), New Mexico Consortium and Los Alamos National Laboratory (Tom Terwilliger's group), Cambridge University (Randy Read's group), Duke University (the Richardsons' group), UTHealth (Matt Baker's group)

APPLICATIONS OF TECHNOLOGY

- Determines three-dimensional (3D) macromolecular structure using crystallography (X-ray, neutron, or electron) or cryo-electron microscopy data.

BENEFITS

- Provides a comprehensive, highly-automated software package for determining macromolecular structure
- Increases the throughput of structure solutions
- Designed for users at all levels, from beginners to experts; Users can easily access the programs through the Phenix graphical user interface (GUI) and/or command line
- Focuses on rapid development and bug fixing
- Provides integration with molecular viewers (Coot and Pymol; coming: ChimeraX)
- Offers convenient ligand handling tools (automated ligand fitting and restraints builders)

BACKGROUND

Knowledge of molecular structures is critical to understanding biological processes and to developing new therapeutics against diseases. Crystallography (X-ray, neutron, and electron) and cryo-electron microscopy are powerful methods for determining 3D macromolecular structures. The overall structure-solution workflow is similar for these techniques, but nuances exist because their reduced experimental data have different properties. Phenix is a comprehensive software package for macromolecular structure determination that handles data acquired using any of these techniques.

TECHNOLOGY OVERVIEW

Researchers at Lawrence Berkeley National Lab and other institutions have developed a comprehensive software package, called Phenix (Python-based Hierarchical ENvironment for Integrated Xtallography), that efficiently and accurately determines macromolecular structures from crystallography and cryo-electron microscopy data.

The Phenix package performs all of the steps required to determine a macromolecular structure. For X-ray diffraction data, these include assessing the data quality, experimental phasing, molecular replacement, building an atomic model, refining and validating the model, and preparing structure deposition into the World-Wide Protein Data Bank. For cryo-electron microscopy data, these include assessing map quality, improving the map, building, refining and validating a model, and preparing structure deposition. The programs developed for each of these steps are highly automated. Thus, Phenix provides rapid structure solutions for both X-ray crystallography and cryo-electron microscopy data—ultimately enabling increased throughput. The Phenix programs are run through a user-friendly interface and/or by command line. The extensive online manual describes GUI and command-line versions of individual programs and includes overviews, tutorials, and FAQs. The Phenix Tutorials YouTube channel provides introductory tutorial videos.


OMICS

basecRAWller: Nanopore Sequencing Base Calling Software

2017-075

Marcus Stoiber, James Brown

APPLICATIONS OF TECHNOLOGY

- Nanopore sequencing (i.e. significantly more accurate base caller for nanopore data)

BENEFITS

- Streaming base calling
- Base calling from information-rich raw signal

BACKGROUND

Nanopore applications require base calling. Current nanopore base calling software applications employ the Viterbi algorithm, which requires the whole sequence to employ the complete base calling procedure and thus precludes a natural streaming base calling procedure.

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab developed the basecRAWller software, the first base caller for nanopore data that calls bases directly from raw data.

In comparison to other sequencing technologies, the basecRAWller algorithm has the powerful ability to perform truly streaming base calling as signal is received from the sequencer. The other major advantage is the prediction of bases from raw signal, which contains much richer information than the segmented chunks that current algorithms employ.

As a result, basecRAWller leads to base calls of much greater accuracy.

Metabolite, Annotation, and Gene Integration System (MAGI) and Software

2017-063, 2017-105

Benjamin Bowen, Markus de Raad, Onur Erbilgin, Trent R Northen, Oliver Reubel

APPLICATIONS OF TECHNOLOGY

- Genomics research

BACKGROUND

Metabolomics has been used for obtaining direct measures of metabolic activities from diverse biological systems. However, metabolomics can be limited by ambiguous metabolite identifications. Furthermore, interpretation can be limited by incomplete and inaccurate genome-based predictions of enzyme activities (e.g., gene annotations). In addition, some genes may be poorly annotated. Thus, the understanding of metabolism, such as microbial metabolism, is limited.

TECHNOLOGY OVERVIEW

A team of Lawrence Berkeley National Laboratory researchers including Trent Northen, Benjamin Bowen, Oliver Reubel, and Markus De Raad has developed technologies for associating metabolites with genes that enable scoring and curating compound identities based on their biological relevance, and/or using compound identities from those tools to connect to genes in their biological samples and potentially formulate hypotheses of gene function. Such results can be used to direct high-throughput biochemical assays to greatly reduce biochemical search space. The Metabolite, Annotation and Gene Integration (MAGI) system is highly relevant to and useful in the fields of genomics, metabolomics, and systems biology. Furthermore, as metabolomics data become more widely available for sequenced organisms, MAGI has the potential to improve the understanding of microbial metabolism, while also providing testable hypotheses for specific biochemical functions.

MAGI v.1.0, a computational software package developed by researchers Onur Erbilgin, Benjamin Bowen, and Oliver Ruebel, efficiently supports the above workflow, integrating experimental metabolomics and genomics data with chemical, biochemical, and genomic data to produce and test hypotheses. In the MAGI workflow, exact chemicals are linked to exact genes via probabilistic relationships between reactions facilitated by a chemical similarity network as well as protein homology searching. MAGI automatically suggests alternative substrates to experimentally test via the chemical similarity network—necessary to determine the specific function of an enzyme—and can execute the experiments in a high throughput manner. MAGI is useful for compound identifications in untargeted metabolomics experiments and annotating genes and genomes, and is most powerful when the two are combined. It is also a major aide for biochemical function discovery, biosynthetic pathway (re) construction, metabolic modeling, and many more aspects of biochemistry. This overcomes previous approaches where researchers need to use several disconnected resources to connect a gene to a compound in a reaction.

OMG: Multi-Omics Data Library for Predictive Modeling

2021-007

Somtirtha Roy, Hector Garcia Martin, Jose Manuel Martin, Tijana Radivojevic

APPLICATIONS OF TECHNOLOGY

- Predictive modeling for bioengineering and synthetic biology

BENEFITS

- Enables accurate testing of biological computational and algorithmic tools
- Provides reliable synthetic biological data

BACKGROUND

Synthetic biology cannot yet fulfill its potential due to the inability to predict the behavior of biological systems. Computational tools can move the field forward by leveraging multiomics data to predict the outcome of bioengineering efforts.

TECHNOLOGY OVERVIEW

The Omics Mock Generator (OMG) library, created by Berkeley Lab scientists, is used to provide synthetic multiomics data for testing computational tools for bioengineering metabolic models. Since experimental multiomics data is expensive to produce, OMG provides a simple and efficient way to produce large amounts of multiomics data. This data is both accessible and also biologically accurate such that it can be used to test algorithms and tools systematically.

Omics Mock Generator works by creating fluxes based on Flux Balance Analysis (FBA) and growth rate maximization, leveraging COBRApy. OMG is compatible with any genome-scale model. In order to obtain proteomics data, it can be assumed that the corresponding protein expression and gene transcription are linearly related to the fluxes, while the amount of metabolite present is assumed to be proportional to the sum of absolute fluxes coming in and out of the metabolite.

Experiment Data Depot (EDD): Biological Experimental Data Storage and Visualization

2016-186

Hector Garcia Martin, Garrett Birkel, William Morrell, Mark Forrer, Teresa Lopez

APPLICATIONS OF TECHNOLOGY

- Biological experiment data storage, sharing, and visualization

BENEFITS

- Provides a novel and convenient way to store a variety of data types, visualize these data, and export them in a standardized fashion for use with predictive algorithms

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab have developed the Experiment Data Depot (EDD), a web-based service for storing and visualizing data collected in biological experiments, including measurements, protocols, and metadata. EDD provides a novel way to keep all actionable data from an experiment in a single place, where this data can be conveniently browsed and quality checked. EDD hosts different types of data (proteomics, metabolomics, HPLC, etc.) so as to provide actionable information. EDD output is standardized into different output types that can be immediately used for modeling the systems.

EDD provides a standardized description of experiments and significantly facilitates the Test and Learn phases in the Design-Build-Test-Learn (DBTL) cycle. Its output types can be immediately used for modeling systems and it interacts easily with other services such as JBEI's ICE (the Inventory of Composable Elements) and Arrowland (multiomics visualization tool).

DelPlasmid: Finding Plasmids with Deep Learning and Machine Learning

2019-037

William Andreopoulos, Jan Belewski

APPLICATIONS OF TECHNOLOGY

- Identification of plasmids in assembled bacterial genomes

BENEFITS

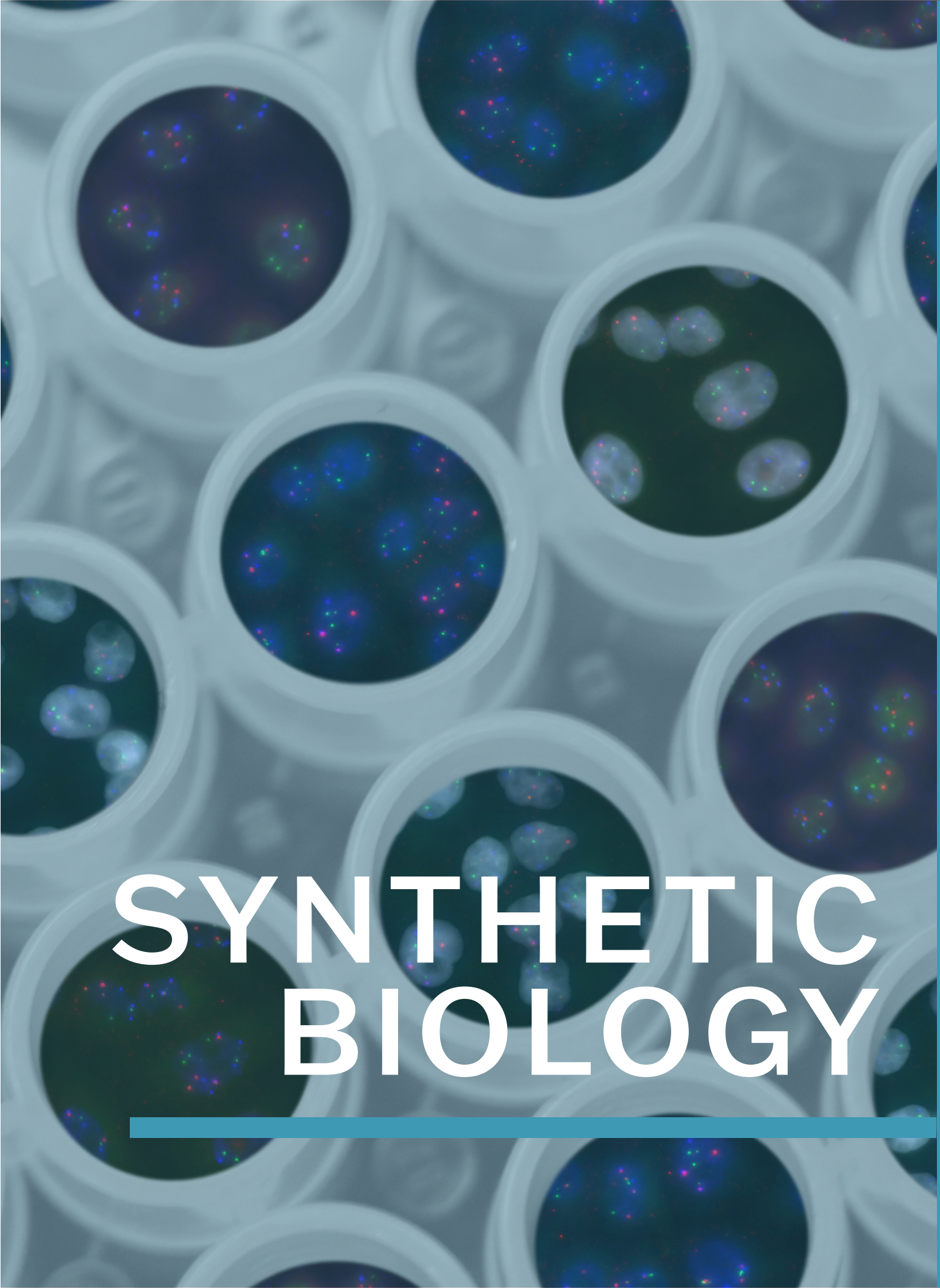
- Recovery of plasmids from assembled bacterial genomes without any prior taxonomical knowledge with a low false positive rate

BACKGROUND

Plasmids are extrachromosomal genetic elements that are an important driver of DNA exchange and genetic innovation in prokaryotes. The success of plasmids has been attributed to their independent replication from host cell chromosomes and their frequent self-transfer. In recent years, the Department of Energy (DOE) has recognized the importance of plasmid DNA to the viability and success of microbial communities.

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab have developed DelPlasmid, a novel software that helps identify plasmids in assembled bacterial genomes. DelPlasmid is Long Short-Term Memory (LSTM)-based on a deep learning model that takes as input a combination of assembled sequences and extracted features to identify bacterial plasmids. This model was trained on high-quality plasmid sequences from the ACLAME database and the NCBI Refseq.microbial dataset. The tool achieved an AUC-ROC of 91% on a 5-fold cross-validation.



SYNTHETIC BIOLOGY

DIVA: DNA Design, Implementation, and Validation Automation Software

2021-102

Hector Plahar, Lisa Simirenko, Manjiri Tapaswi, Steve Lane

APPLICATIONS OF TECHNOLOGY

- DNA design and construction
- Synthetic biology, engineering biology, efficiency metric tracking

BENEFITS

- Provides a collaborative web-based graphical user interface supporting and tracking DNA design and construction
- Interfaces with other related enabling software to offer additional functionality and reduce friction for the user

BACKGROUND

DIVA is a new graphical user interface for the j5 DNA assembly design software. It leverages other software tools including ICE (Inventory of Composable Elements) to source and deposit DNA sequences, OpenVectorEditor to visualize DNA sequences, BOOST for protein sequence back translation (and other functionality), and BLiSS for biosecurity screening.

TECHNOLOGY OVERVIEW

Researchers at Berkeley and Sandia National Labs have designed a web-based software platform known as DIVA (Design, Implementation, and Validation Automation), which improves the operational efficiency of companies and research institutions working on DNA design and construction.

DIVA enables users to collaboratively and visually design DNA constructs and submit the designs to internal resources or outside vendors for construction at scale. Users design DNA with a web-based graphical user interface, submit their designs to a central queue, and after internal or external review, receive their sequence-verified clonal constructs. Users can track the construction and subsequent sequence validation processes using DIVA. Completely independent DNA construction tasks can be aggregated into the same multi-well plates and pursued in parallel because the DIVA DNA construction methods are sequence-agnostic and standardized, i.e. they use the same enzymatic master mixes and reaction conditions.

Since DIVA captures in a central database all DNA construction success and failure rates, including at which stage a given clonal construction procedure failed, the data can be leveraged to refine and track the efficiency of the DNA assembly design process.

BOOST (Build Optimization Software Tools) for DNA Sequence Design

2016-107

Ian Blaby, Gautham Mani, Ernst Oberortner, Jan-Fang Cheng, Nathan J. Hillson, Samuel Deutsch

APPLICATIONS OF TECHNOLOGY

- Synthetic biology
- Genomics
- Gene characterization
- Drug discovery

BENEFITS

- Accelerates gene discovery and characterization toward practical applications
- Significantly reduces cost and turnaround time of DNA sequence synthesis
- Automatically detects and redesigns difficult sequences for DNA synthesis with higher success rates, eliminating trial-and-error processes
- Converts files among common synthetic biology software formats

BACKGROUND

Synthetic biology enables the design and synthesis of DNA sequences that are difficult to find in nature, as well as sequences that are not known to occur naturally, with potential applications in diverse fields, such as fuels, chemical production, and cellular engineering, among others. However, not every DNA sequence can be readily manufactured due to current limitations of biological design software tools and constraints of DNA synthesis.

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab and the U.S. Department of Energy Joint Genome Institute have developed a suite of build-optimization software tools (BOOST) to streamline the design-build transition in synthetic biology engineering workflows. The BOOST library offers a wide range of capabilities, including:

- Reverse translation and codon juggling
- Detection and resolution of constraint violations
- Polishing of individual sequences
- Sequence partitioning
- The detection of constraint violations preempts the need for sequence redesign by users. In addition, by optimizing the design of DNA sequences, the BOOST library can significantly reduce the cost and turnaround time of DNA synthesis, compared to commercial DNA design software tools.

SynTrack: DNA Assembly Workflow Management

2017-023

Xianwei (John) Meng, Lisa Simirenko

APPLICATIONS OF TECHNOLOGY

- Management and tracking of DNA synthesis and assembly operations
- Integration of quality control outcomes and the status of construct deliverables for external users

BENEFITS

System with a host of functions to track the DNA build process:

- Generating Echo pooling instructions based on plate maps
- Bulk updating of colony or PCR amplification information, fusion PCR, and chewback results
- Updating with quality assurance and quality control outcomes with .csv and .xlsx template files
- Re-work assembly workflow enabled before and after sequencing validation

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab have developed SynTrack, a biological computer-aided manufacturing (bioCAM) platform that manages the complexity of synthetic biology workflows and tracks the production of synthetic DNA constructs to maximize efficiency.

SynTrack is integrated with laboratory automation systems by enabling task-based workflows to offer complete DNA construct management, assembly pipeline control, and plate and well content capturing. Functions include:

- Data management of the hierarchical DNA assembly of constructs
- Tracking the usage of the various parts and transitory DNA fragments
- Streamlining DNA construct build processes via workflow template to improve turnaround time
- Flexible handling of the various final DNA constructs from different process tasks
- Generating pipetting instruction to leverage liquid handling robotic platforms
- Genvering vendor DNA fragment order and sample receiving processes
- Management of well plate layouts and associated quantity data

ART: A Machine Learning Automated Recommendation Tool for Guiding Synthetic Biology

2020-011

Brian Foster, Jeff L Froula, Edward Kirton, Angela Kollmer, Mario Melara, Georg Rath, Kelly Rowland, Seung Jin Sul, Stephan Trong

APPLICATIONS OF TECHNOLOGY

- Companies performing synthetic biology/metabolic engineering

BENEFITS

- Demonstrated high predictive accuracy
- Shorter strain development times

BACKGROUND

Large, complex multi-compute-node workflows are not easily maintained, modified, or run on multiple computing resources. This is true for the Berkeley Lab's Joint Genome Institute (JGI) as well as other organizations that run large and complex calculations. The JGI Analysis Workflow Service (JAWS) aims to improve the reusability and robustness of bioinformatic and other workflows in evolving and diverse HPC, large cluster, and cloud environments.

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab's Joint BioEnergy Institute and the Agile BioFoundry have developed a tool that uses machine learning to make further advancements in the field of synthetic biology. The patent-pending Automated Recommendation Tool (ART) uses probabilistic modeling techniques to guide metabolic engineering systematically without requiring a full mechanistic understanding of the biological system. Using sampling-based optimization, ART provides a set of recommended strains to be built in the next engineering cycle, alongside probabilistic predictions of their production levels. ART is built around a unique uncertainty quantification approach and has been demonstrated to have high predictive accuracy. Using ART improved tryptophan titer and productivity by up to 74% and 43%, respectively, compared to the best designs used for algorithm training.

ClusterCAD: Computational Platform for Design of Chimeric Polyketide Synthase (PKS) Enzymes

2017-101

Clara H Eng, Tyler W H Backman, Constance B Bailey, Christophe Magna, Hector Garcia Martin, Leonard Katz, Pierre Baldi, Jay D Keasling

APPLICATIONS OF TECHNOLOGY

- Drug development
- Bioproducts
- Chemical Engineering

BENEFITS

- Saves time in developing novel or drop-in molecules using type I PKS
- Easily identifies polyketide synthase (PKS) modules based either on amino acid sequence or on the chemical structure of the cognate polyketide intermediate
- More reliable construction of functional PKS chimeras

BACKGROUND

Type I modular polyketide synthases (PKS) have a unique modular structure in which the product of each module is determined by the catalytic domains that comprise each module. This modular nature suggests that their biosynthetic power can be harnessed for combinatorial biosynthesis. Previous work has demonstrated that it is possible to construct functional chimeric PKSs by exchanging catalytic domains between heterologous PKS modules. However, limitations in the theoretical understanding of the complex protein-protein interactions that govern the fold and function of natural and engineered PKSs mean that identifying strategies to reliably design functional PKS chimeras remains an open research problem.

TECHNOLOGY OVERVIEW

Researchers at Berkeley Lab have developed a computational platform, known as ClusterCAD, that facilitates the informed design of chimeric type I modular PKSs in order to achieve the microbial production of novel drug analogs and industrially relevant small molecules. These type I modular PKS clusters are contained in the ClusterCAD database and are linked to the corresponding MiBiG database and NCBI Nucleotide database entries. It provides chemical structures with stereochemistry for the intermediates generated by each PKS module, as well as sequence- and structure-based search tools that allow users to identify modules based either on amino acid sequence or on the chemical structure of the cognate polyketide intermediate. This facilitates the process behind chimeric PKS design, as the nature of protein-protein interactions are very complex.

ClusterCAD provides the first database of PKS biosynthetic clusters that contains predicted chemical structures for the polyketide intermediates generated by each module, as well as the predicted relatively solvent accessibility and secondary structure of each subunit. With search tools to query the database for modules based on either amino acid sequence, or on the chemical structure of the cognate polyketide intermediate, ClusterCAD provides the first computational platform developed for computer-aided design of chimeric PKSs.

ICE: Inventory of Composable Elements (ICE), JBEI Registry of Biological Parts, Web of Registries (WoRS)

2639, 2020-139

Hector Plahar, Steve Lane, Wiliam Morrel

APPLICATIONS OF TECHNOLOGY

- Search, storage, and management of databases of biological parts and associated sequences

BENEFITS

- Efficient sourcing and management of data
- Facilitates data sharing and exchange

TECHNOLOGY OVERVIEW

ICE and Web of Registries
The Inventory of Composable Elements (ICE) is an open source web based registry platform for synthetic biological parts with support for microbial strains, plasmids, Arabidopsis seeds, and generic parts. The integration of a plasmid/vector editor enables the visualization and vector manipulation of any associated sequences. It can also be used by laboratories to track and search their constructs. It has an implementation of the Web of Registries concept that enables multiple standalone ICE instances to interconnect and form a distributed parts database.

Bioparts Search Portal
The Bioparts Search Portal is a web based application that enables users to search for publicly available biological parts using keywords or sequence fragments. For the initial version (1.0) of the application, Bioparts targets 10 sources of biological part data: the GenBank NIH genetic sequence database (https://www.ncbi.nlm.nih.gov/genbank/), the iGem parts registry (parts.igem.org), the Addgene plasmid repository (https://www.addgene.org/), and seven Inventories of Composable Elements (ACS Synbio, JGI, JBEI, JBEI Public, ABF, SynBerc, ABF Public). Bioparts has built-in automated web scrapers which extract data from sources that do not have a public or well-defined application programming interface (API). They extract as much public data as they can find and create a searchable index to speed up searches. Included in the indexed information is the source of the information. Additionally, the portal offers a REST API that enables third-party applications and tools to access the portal's functionality programmatically.

HIGH PERFORMANCE COMPUTING

JAWS (Joint Genome Institute Analysis Workflow Service) for Complex Computational Pipelines on Multiple Compute Resources

2020-090

Brian Foster, Jeff L Froula, Edward Kirton, Angela Kollmer, Mario Melara, Georg Rath, Kelly Rowland, Seung Jin Sul, Stephan Trong

APPLICATIONS OF TECHNOLOGY

- Running large, complex computational workflows, and pipelines (e.g., Broad Institute's Cromwell)
- Data analysis workflows in High Performance Computing (HPC), large cluster, and cloud environments

BENEFITS

- Improves reusability and robustness of workflows/pipelines
- Allows for easier maintenance and modification
- Enables distribution of workflows across different computing facilities

BACKGROUND

Large, complex multi-compute-node workflows are not easily maintained, modified, or run on multiple computing resources. This is true for the Berkeley Lab's Joint Genome Institute (JGI) as well as other organizations that run large and complex calculations. The JGI Analysis Workflow Service (JAWS) aims to improve the reusability and robustness of bioinformatic and other workflows in evolving and diverse HPC, large cluster, and cloud environments.

TECHNOLOGY OVERVIEW

The Joint Genome Institute (JGI) at Berkeley Lab has developed JAWS as a framework to run complex computational workflows at one or more compute sites (e.g. grid-clusters, cloud resources). JAWS moves data and code to user-specified compute resources, executes computation, and returns results. JAWS leverages Globus, developed at the University of Chicago, which implements GridFTP for high-throughput transfer of files (e.g. workflows' input data, output results) between the client server and computation sites (i.e. compute clusters). JAWS utilizes Cromwell, developed at the Broad Institute, to execute workflows written in either Workflow Description Language (WDL) or Common Workflow Language (CWL). Additionally, JAWS supports containers (compatible with Docker, Shifter, and Singularity) to ensure use of well-defined compute environments and workflow task codebases. To run the compute tasks, JAWS provides the JGI Task Manager (JTM), a compute task framing tool that manages workers on compute nodes and executes tasks on them.

DEVELOPER LIST

SPECTRAL IMAGING

BASTet and OMAAT

Oliver Ruebel, Ben Bowen, Tristan de Rond, Markus de Raad

SPECT

Rostylav Buchko, Grant T Gullberg, Bryan W Reutter, Yuval Zelnik

LABELIT

Nicholas K. Sauter

PHENIX

Paul Adams	*Lawrence Berkeley National Laboratory*
Pavel Afonine	*Lawrence Berkeley National Laboratory*
Matt Baker	*UT Health*
Vincen Chen	*Duke University*
Tristan Croll	*University of Cambridge*
Li-Wei Hung	*Los Alamos National Laboratory*
Corey Hryc	*UT Health*
Dorothee Liebschner	*Lawrence Berkeley National Laboratory*
Airlie McCoy	*University of Cambridge*
Claudia Millan Nebot	*University of Cambridge*
Nigel Moriarty	*Lawrence Berkeley National Laboratory*
Rob Oeffner	*University of Cambridge*
Billy Poon	*Lawrence Berkeley National Laboratory*
Randy Read	*University of Cambridge*
Jane Richardson	*Duke University*
Dave Richardson	*Duke University*
Christopher Schlicksup	*Lawrence Berkeley National Laboratory*
Oleg Sobolev	*Lawrence Berkeley National Laboratory*
Tom Terwilliger	*Los Alamos National Laboratory*
Christopher Williams	*Duke University*
Gabor Bunkóczi	*University of Cambridge*
Youval Dar	*Lawrence Berkeley National Laboratory*
Ian Davis	*Duke University*
Lindsay Deis	*Duke University*
Nat Echols	*Lawrence Berkeley National Laboratory*
Richard Gildea	*Lawrence Berkeley National Laboratory*
Kreshna Gopal	*Texas A&M University*
Ralf Grosse-Kunstleve	*Lawrence Berkeley National Laboratory*
Hamsapriye	*University of Cambridge*
Kaushik Hatti	*University of Cambridge*
Jeff Headd	*Lawrence Berkeley National Laboratory*
Bradley Hintze	*Duke University*
Bob Immormino	*Duke University*
Tom Ioerger	*Texas A&M University*
Swati Jain	*Duke University*
Lalji Kanbi	*Texas A&M University*
Gary Kapral	*Duke University*
Erik McKee	*Texas A&M University*

Michael Prisant	*Duke University*
Reetal Pai	*Texas A&M University*
Thiru Radhakannan	*Los Alamos National Laboratory*
Ian Rees	*Lawrence Berkeley National Laboratory*
Tod Romo	*Texas A&M University*
James Sacchettini	*Texas A&M University*
Massimo Sammito	*University of Cambridge*
Duncan Stockwell	*University of Cambridge*
Jacob Smith	*Texas A&M University*
Laurent Storoni	*University of Cambridge*
Alexandre Urzhumtsev	*Institut de Génétique et de Biologie Moléculaire et Cellulaire*
Lizbeth Videau	*Duke University*
Peter Zwart	*Lawrence Berkeley National Laboratory*

OMICS

basecRAWller

Marcus Stoiber, James Brown

MAGI

Benjamin Bowen, Markus de Raad, Onur Erbilgin, Trent R Northen, Oliver Reubel

OMG

Somtirtha Roy, Hector Garcia Martin, Jose Manuel Martin, Tijana Radivojevic

EDD

Hector Garcia Martin, Garrett Birkel, William Morrell, Mark Forrer, Teresa Lopez

DelPlasmid

William Andreopoulos, Jan Belewski

SYNTHETIC BIOLOGY

DIVA

Hector Plahar, Lisa Simirenko, Manjiri Tapaswi, Steve Lane

BOOST

Ian Blaby, Gautham Mani, Ernst Oberortner, Jan-Fang Cheng, Nathan J. Hillson, Samuel Deutsch

SynTrack

Xianwei (John) Meng, Lisa Simirenko

ART

Brian Foster, Jeff L Froula, Edward Kirton, Angela Kollmer, Mario Melara, Georg Rath, Kelly Rowland, Seung Jin Sul, Stephan Trong

ClusterCAD

Clara H Eng, Tyler W H Backman, Constance B Bailey, Christophe Magna, Hector Garcia Martin, Leonard Katz, Pierre Baldi, Jay D Keasling

ICE

Hector Plahar, Steve Lane, Wiliam Morrel

HIGH PERFORMANCE COMPUTING

JAWS

Brian Foster, Jeff L Froula, Edward Kirton, Angela Kollmer, Mario Melara, Georg Rath, Kelly Rowland, Seung Jin Sul, Stephan Trong

BERKELEY LAB

Bringing Science Solutions to the World

Founded in 1931 on the belief that the biggest scientific challenges are best addressed by teams, Lawrence Berkeley National Laboratory and its scientists have been recognized with 14 Nobel Prizes. Today, Berkeley Lab researchers develop sustainable energy and environmental solutions, create useful new materials, advance the frontiers of computing, and probe the mysteries of life, matter, and the universe. Scientists from around the world rely on the Lab's facilities for their own discovery science. Berkeley Lab is a multiprogram national laboratory, managed by the University of California for the U.S. Department of Energy's Office of Science.