Report of ASCAC Subcommittee for DOE-NCI Collaborative Program

Tony Hey

September 2022

Charge Letter to ASCAC

- The 2016 Memorandum of Understanding for the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) has been renewed for an additional five-year period.
- ASCAC was requested to form a working group to review the activities under this collaboration to:
    - provide advice to the Office of Science regarding new opportunities that might contribute significantly to these efforts
    - identify any major challenges that are preventing the efforts from delivering on their potential
    - provide recommendations for how the Office of Science might address these challenges
- The working group should report findings through the ASCAC annually to identify significant opportunities and challenges in a timely manner.
- The ASCAC Chair will transmit the working group's findings in the form of a letter report to the Director of the Office of Science after the letter report is accepted by the full committee at an open public meeting.

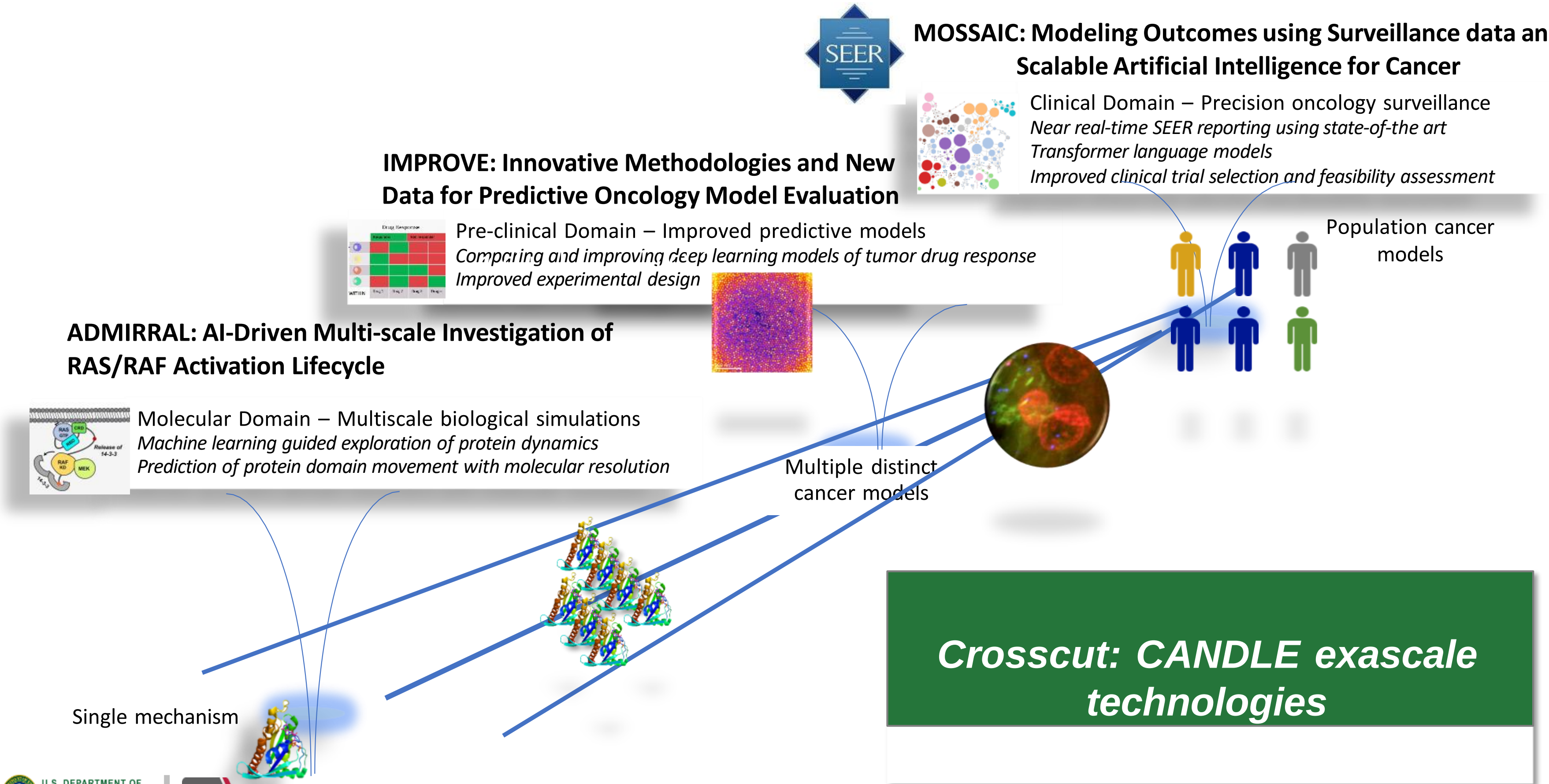
New MOU June 2021:
Collaboration Governance and Oversight

	U.S. DEPARTMENT OF ENERGY Office of Science NCI-DOE ASCAC Subcommittee	NCI-DOE Collaboration Scientific & Technical Advisory Committees	NCI-DOE Collaboration Executive Committee
Number	1	1 per project	1
Member composition	Chair: Tony Hey. Members: 8-12 extramural scientists with expertise across collaboration areas (cancer, biology, advanced computing, data science, etc.)	4-6 scientists per committee with targeted, deep expertise relevant to the assigned project	NCI: Drs. Sharpless, Lowy, Singer DOE: Drs. Binkley & Helland (SC), Dr. Anderson & Ms. Hoang (NNSA)
Member selection	per ASCAC guidance with input from NCI and DOE leadership	by project leads in consultation with Exec Committee	by agency leadership
Meeting Frequency	2 times per year or as determined by Subcommittee chair	Quarterly or as needed	3 times per year
Charge/role	- Assessment of current projects - Assessment of opportunities and challenges - Identification of strategies to address challenges and deliver on opportunities	Project-specific, in-depth scientific and technical guidance and advisement	- Interagency strategic partnership status and relationship health - Overall funding - Program priorities - Implementation of ASCAC recommendations

ASCAC DOE-NCI Subcommittee

- Tony Hey – STFC, ASCAC (Chair)
- Rick Arthur – GE, ASCAC
- Jay Bardhan – PNNL
- Martin Berzins – Utah, ASCAC
- Bill Gropp – UIUC, NCSA
- Satheesh Maheswaran – DLS, UKRI
- Amanda Randles – Duke
- Vadim Backman – Northwestern
- Caroline Chung – MD Anderson
- Susan Gregurick – NIH, ASCAC
- Amie Hwang – USC
- Gordon Mills - OHSU
- Joel Saltz – Stony Brook

JDACS4C Projects: Towards Predictive Oncology



MOSSAIC: Modeling Outcomes using Surveillance data and Scalable Artificial Intelligence for Cancer

Clinical Domain – Precision oncology surveillance

Near real-time SEER reporting using state-of-the art Transformer language models

Improved clinical trial selection and feasibility assessment

IMPROVE: Innovative Methodologies and New Data for Predictive Oncology Model Evaluation

Pre-clinical Domain – Improved predictive models

Comparing and improving deep learning models of tumor drug response

Improved experimental design

ADMIRRAL: AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle

Molecular Domain – Multiscale biological simulations

Machine learning guided exploration of protein dynamics

Prediction of protein domain movement with molecular resolution

Crosscut: CANDLE exascale technologies

U.S. DEPARTMENT OF ENERGY | NIH NATIONAL CANCER INSTITUTE

NCI-DOE Subcommittee Activity in 2022

Meetings

- Introductory meeting with PIs – January 18th
- MOSSAIC Review – March 28th
- ADMIRRAL Review – July 28th
- IMPROVE and CANDLE Reviews – September 16th

Process

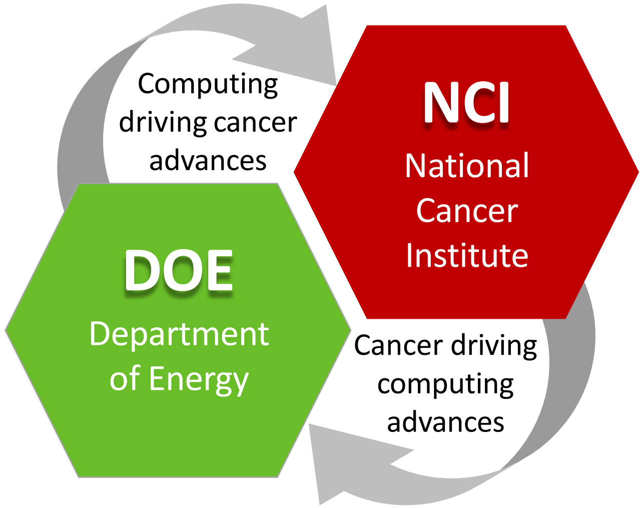
- Two subcommittee members assigned as leads for project reviews and to summarize the subcommittee's conclusions

Project Reviews

September 30, 2022

MOSSAIC: **M**odeling **O**utcomes using **S**urveillance data and **S**calable **AI** for **C**ancer

DOE-NCI partnership to advance exascale development through cancer research


Computing driving cancer advances
NCI
National Cancer Institute
DOE
Department of Energy
Cancer driving computing advances

Lynne Penberthy
National Cancer Institute

Georgia Tourassi
Oak Ridge National Laboratory

March 28, 2022

Presented to:
ASCAC DOE-NCI Subcommittee

This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health. This work was performed under the auspices of the U.S. Department of Energy by Argonne National Laboratory under Contract DE-AC02-06-CH11357, Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344, Los Alamos National Laboratory under Contract DE-AC5206NA25396, and Oak Ridge National Laboratory under Contract DE-AC05-00OR22725. [Also adding in the appropriate review and release number]

MOSSAIC Project

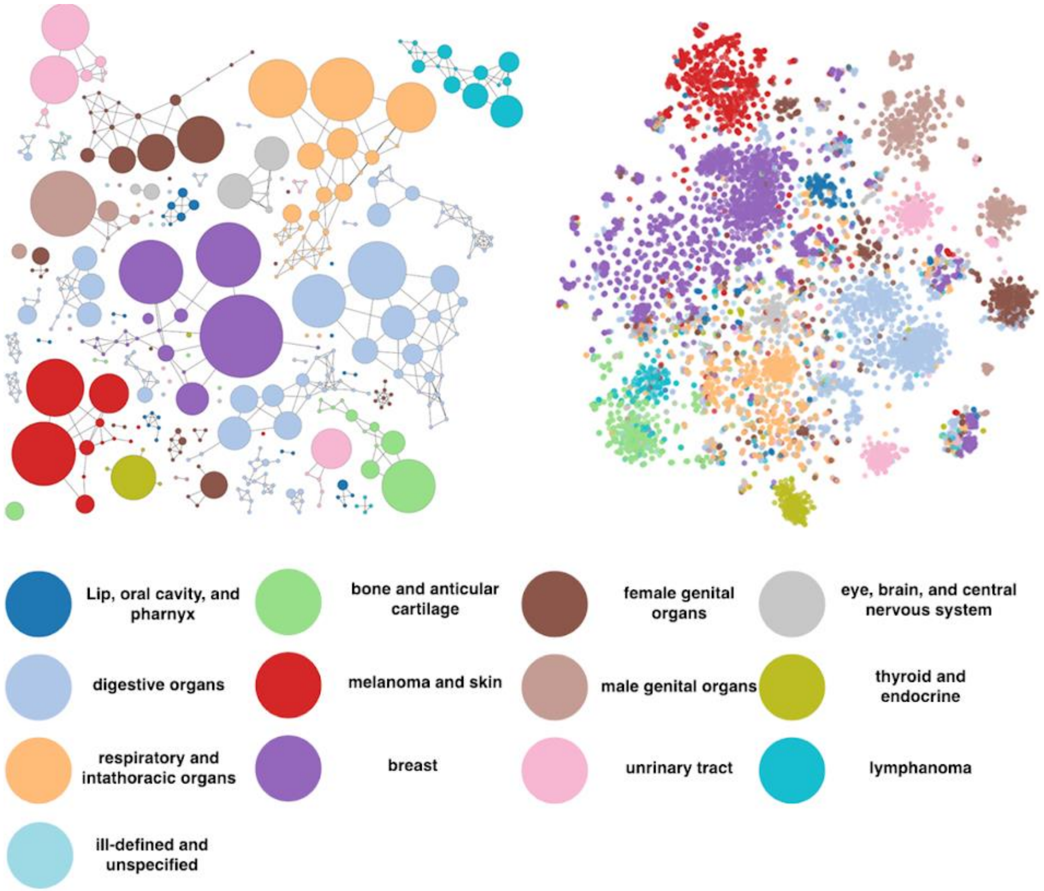
Goal: Modernize national cancer surveillance program (NCI’s SEER program) by developing and deploying scalable deep learning solutions

Data:

- 7 SEER cancer registries
- 1M unique cancer patients
- 3.5M cancer pathology reports
- Annotations for key data elements such as site, histology, behavior, etc.

Key Activities:

1. Developing large-scale, state-of-the-art Transformer language models for clinical information extraction
2. Building new capabilities for biomarker and recurrence detection
3. Pushing novel research in uncertainty quantification
4. Conducting lab studies to evaluate the performance of models in real-world cancer registries
5. Enabling large-scale Transformer training on LCF systems



MOSSAIC Project: Findings

- Extracts detailed and nuanced clinical information encompassing cancer diagnosis, pathology, biomarkers, initial and follow-on treatments, metastases and recurrence
- Automates cancer information abstraction for NCI SEER cancer registries - carrying out coding that would otherwise require human abstractors
- Labels pathology reports to indicate SEER eligibility
- Deployed in 11 SEER and one non-SEER registries
- Extremely high positive predictive values with uncertainty quantification
  - Selects 17% of pathology reports and extracts information with 98% accuracy
- Uses novel uncertainty quantification methods

MOSSAIC Project: Comments

- Outstanding work in creating demonstrably scalable NLP clinical data extraction system
- Carried out extensive work on benchmarking and model development
  - The currently deployed model is a Multi-Task Hierarchical Self Attention Network (HiSAN)
  - the group has also been actively exploring transformer based models
- API has been developed and deployment ongoing in non-SEER registries
  - Any registry using the SEER registry software SEER*DMS can request production deployment
- Collaboration with CDC to develop a privacy preserving API
- Leverages Oak Ridge National Laboratory HIPAA compliant HPC environment tools
- Uses OLCF developed data transfer protocols and the CITADEL framework for secure Summit access and scalable analyses
  - Extensive use of Summit (300K Summit node hours) and ongoing migration to Frontier.
- Dissemination and Academic Productivity
  - ~40 articles and conference papers
  - three public codes released via CANDLE on Github
  - disseminated trained AI models to SEER and stakeholders
  - conducted NLP hackathons through the Summit environment

MOSSAIC Project: Recommendations

1. Committee lauds the team's interest in the use of multi-modal data analyses encompassing pathology and radiology reports. We encourage exploring paths forward to expanding this effort to include discrete data elements, other clinical notes along with judicious use of radiology and pathology image data.

2. The software and methods developed through this project can be applied to cancer clinical data extraction in many additional translational research settings. The project specifications called for modest initial recall with very high positive predictive values. Most research scenarios would benefit from extracting clinical data from a much larger fraction of reports even at the cost of lower positive predictive value. It seems likely that the uncertainty quantification methods could be easily adapted to provide a rough assessment of data extraction confidence.

3. The methods can also be extended to address extraction of a much broader range of cancer and non-cancer related clinical phenotypes. In this context, it would be useful to know how the performance characteristics of the MOSSAIC pipelines compare with pipelines created by other clinical informatics groups.

JDACS4C and the ADMIRRAL Project

Dwight V. Nissley

Frederick National Laboratory for Cancer Research
US National Institutes for Health

Frederick H. Streitz

Center for Forecasting and Outbreak Analytics, US CDC
Lawrence Livermore National Laboratory, US DOE

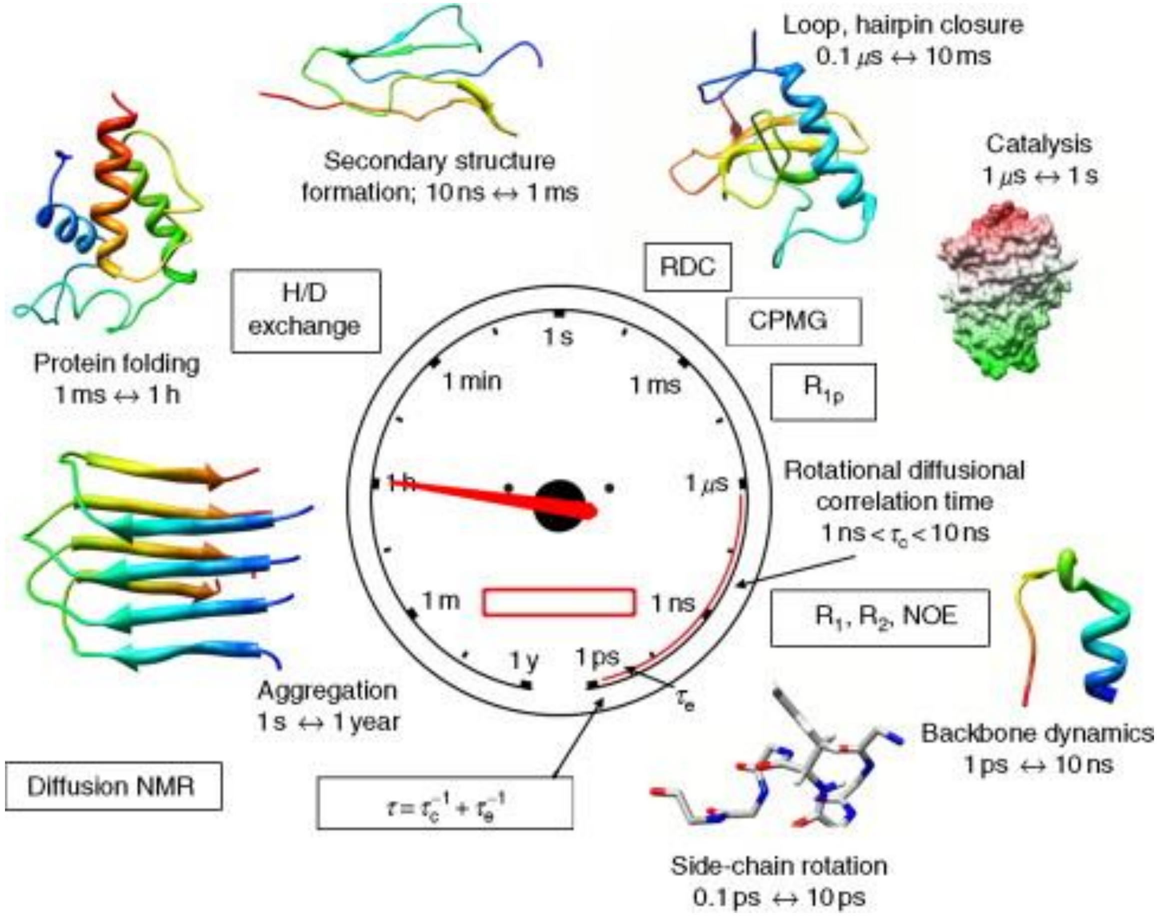
ADMIRRAL Overview

ASCAC Subcommittee

July 28, 2022

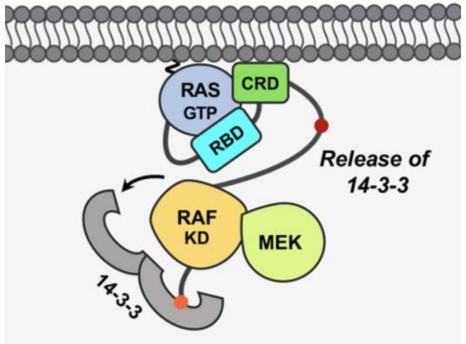
ADMIRRAL: AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle

Co-PIs: Fred Streitz (LLNL), Dwight Nissley (FNL)



Secondary structure formation; 10 ns ↔ 1 ms
Loop, hairpin closure 0.1 μs ↔ 10 ms
Catalysis 1 μs ↔ 1 s
Protein folding 1 ms ↔ 1 h
H/D exchange
RDC
CPMG
$R_{1\rho}$
1 s
1 min
1 ms
1 h
1 μs
Rotational diffusional correlation time $1\,ns < \tau_c < 10\,ns$
1 m
1 ns
R_1, R_2, NOE
1 y
1 ps
τ_e
Aggregation 1 s ↔ 1 year
Backbone dynamics 1 ps ↔ 10 ns
Diffusion NMR
$\tau = \tau_c^{-1} + \tau_e^{-1}$
Side-chain rotation 0.1 ps ↔ 10 ps

- New MuMMI framework that reaches beyond existing model to enable ML-guided exploration of protein dynamics
- Hypothesize long-time step conformation changes that are validated through a suite of fine-grained simulations
- **Intractable biological challenge – predict protein domain movement with molecular resolution**



ADMIRRAL Project: Background

Mis-regulated RAS signaling can result in cancer

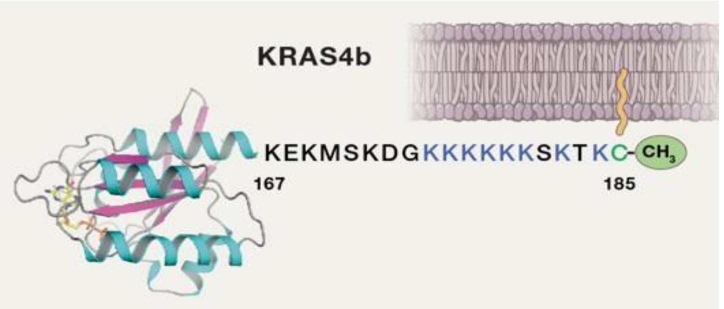
93% of pancreatic cancer

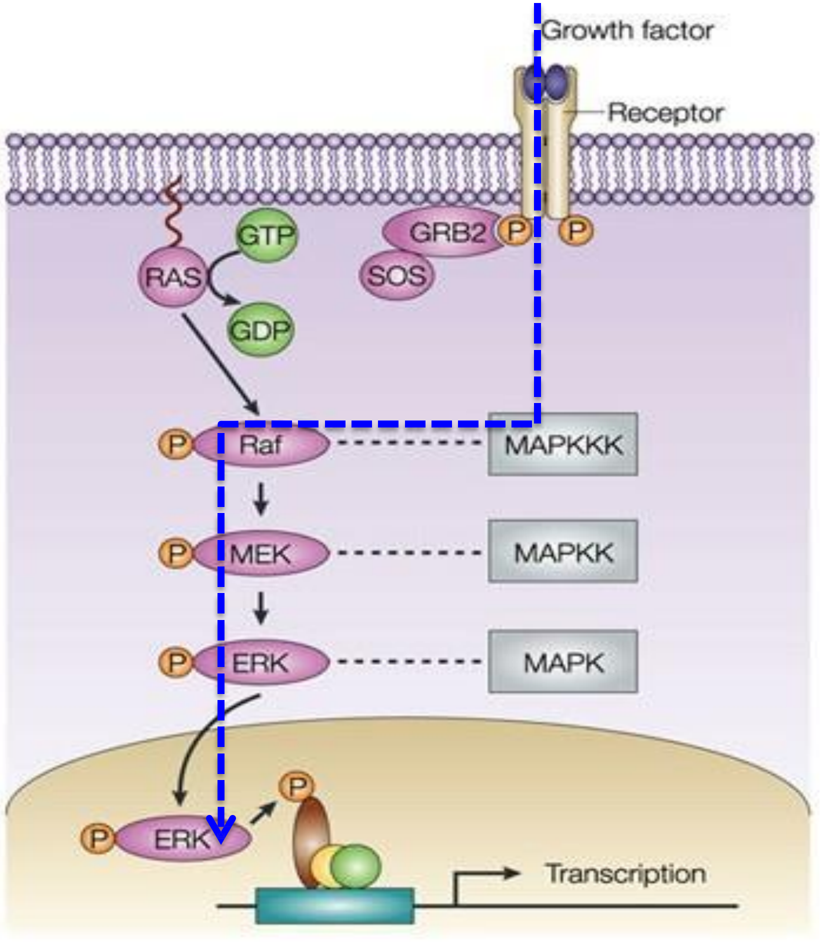
42% of colorectal cancer

33% of lung cancers

1 million deaths/year world-wide

Effective inhibitors needed





Nature Reviews | Molecular Cell Biology

- Pathway transmits signals to the nucleus

- RAS is a switch

  oncogenic RAS is “always on”

- RAS localizes to the plasma membrane

- RAS binds effectors (RAF) to activate growth

ADMIRRAL Project: Findings

- ADMIRRAL is an important study investigating the dynamics of RAS-RAF interactions

- The team has impressively combined the experimental approaches driven from Frederick National Laboratory with the high-performance computing expertise at Livermore National Laboratory

- They have produced strong contributions in terms of new infrastructure for using machine learning to support modeling across spatial and temporal scales

ADMIRRAL Project: Comments

- Understanding the dynamics of RAS-RAF interactions across scales is significant due to the high rate of cancers associated with RAS-RAF mutations.
- The multiscale methods developed in this work lay the groundwork for assessing a wider range of molecular machines.
- The team has delivered advances in multiscale modeling, particularly the data-driven workflow.
- The project team has done an impressive job combining expertise from the DOE team at LLNL and the experimental team at NCI's Frederick National Laboratory for Cancer Research.
- The number of different experimental techniques, and the number of observables that could possibly be computed, seem to present a large number of as-yet-untapped opportunities to test the models.

ADMIRRAL Project: Recommendations

1. Extension of in vitro analysis to explore other forms of RAS and KRA mutations
2. Take into account other proteins that are present and assess their impact
3. Test in intact cells
4. With the initial pilot in place, put emphasis on generalizability of the infrastructure for application to other molecular machiens
5. Address how the models will be tested with experimental studies
6. Increase engagement with broader research communities

IMPROVE: *Innovative Methodologies and New Data for Predictive Oncology Model Evaluation*

[STRUCTURE] + [PROTOCOL] + [DATA] ⟹ [MODEL]

Aim-1: (IMPROVING future models) by developing a scalable, generalizable framework

- to compare deep learning models from the cancer drug modeling community, and
- identify model attributes that contribute to a models' prediction performance

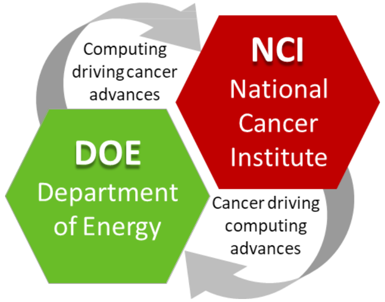
Aim-2: (IMPROVING model performance) by developing protocols

- for identifying and generating targeted data explicitly aimed at strengthening drug response model predictive capabilities, and
- to organize and publish a set of standard datasets, test and validation datasets, and protocols for comparison and evaluation.

Rick Stevens
Argonne National Laboratory
University of Chicago

Jeff Hildesheim
Division of Cancer Biology
National Cancer Institute

Ryan Weil
Frederick National Laboratory for Cancer Research


Computing driving cancer advances
NCI
National Cancer Institute
DOE
Department of Energy
Cancer driving computing advances

IMPROVE Project: Findings

- Curated References: Cataloged over 120 machine learning drug response prediction models, assessed and automated (via CANDLE) 16 of these for evaluation against benchmark drug screening datasets for predicting drug response and gene expression

- Implemented APIs for generalized analysis across studies and datasets, including prediction performance, learning curve analysis, usefulness, and influence of intrinsic noise on prediction

- Engaged vendors (via RFI) for high-throughput screening on patient-derived cancer models to perform feasibility studies, evaluate correlation between predictions and true response values

- Co-evolved capabilities with CANDLE project for automation-driven productivity for cross-model evaluation, sensitivity analysis, and generalizability

- Published validated models and data to the cancer research community, adhering to FAIR principles

IMPROVE Project: Comments

- Well-structured project focusing on achievable deliverables aligned with both DOE and NCI missions.
- Even greater value by advancing toward the goal for a generalizable framework to compare and improve data-derived models beyond biomedical problems to other application science.
- Challenges remain in lengthy process for model validation and assessing/addressing defects in data quality and insufficiency in training sets (including bias, inconsistency/conflict, and incompleteness).
- Unlike Pilot-1, IMPROVE engages the cancer modeling community to promote buy-in and collaboration.
- Consider exploring novel data/strategies for insights into AI-based modeling of cancer, such as engaging innovative cancer communities observing microenvironments and plasticity.

IMPROVE Project: Recommendations

1. Focus to rigorously prove capability in the context of a small number of selected cancers to show success of the idea of the framework and new data, then generalize by widening assessment context

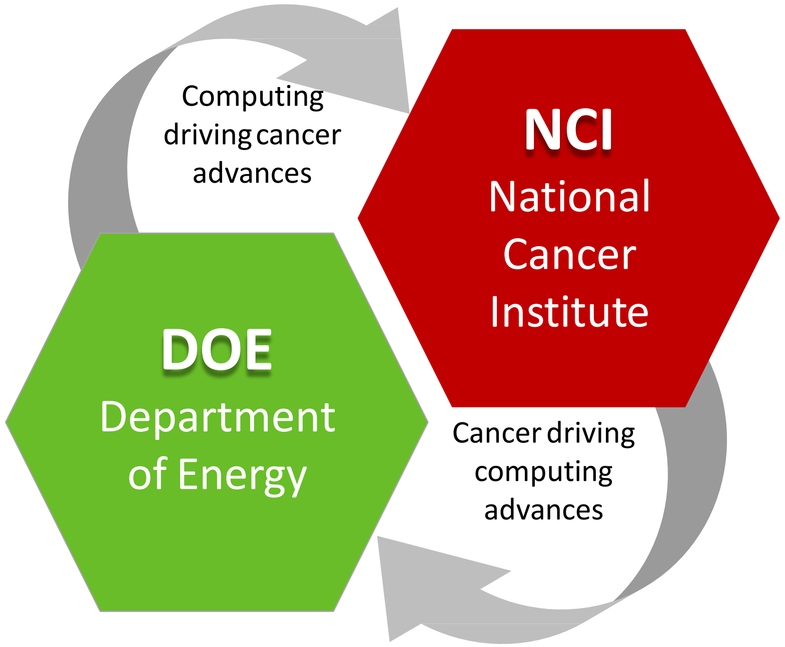
2. Engagement across domains in step with methodical expansion of generalizability to discover and leverage synergies and approaches to common challenges: from cancer-to-cancer to non-cancer diseases, to non-biomedical contexts such as climate studies or materials properties and behaviors

3. Continue community engagement with researchers studying early tumor development, an opportunity for keen insight into the types of data that may strengthen predictive models

4. Consider developing metrics to baseline and form a measure for assessment and comparison inclusive of factors such as productivity, completeness, correctness, and consistency

CANcer Distributed Learning Environment (CANDLE)

DOE-NCI partnership to advance exascale development through cancer research



Computing driving cancer advances

NCI
National Cancer Institute

DOE
Department of Energy

Cancer driving computing advances

Rick Stevens
Argonne National Laboratory
University of Chicago

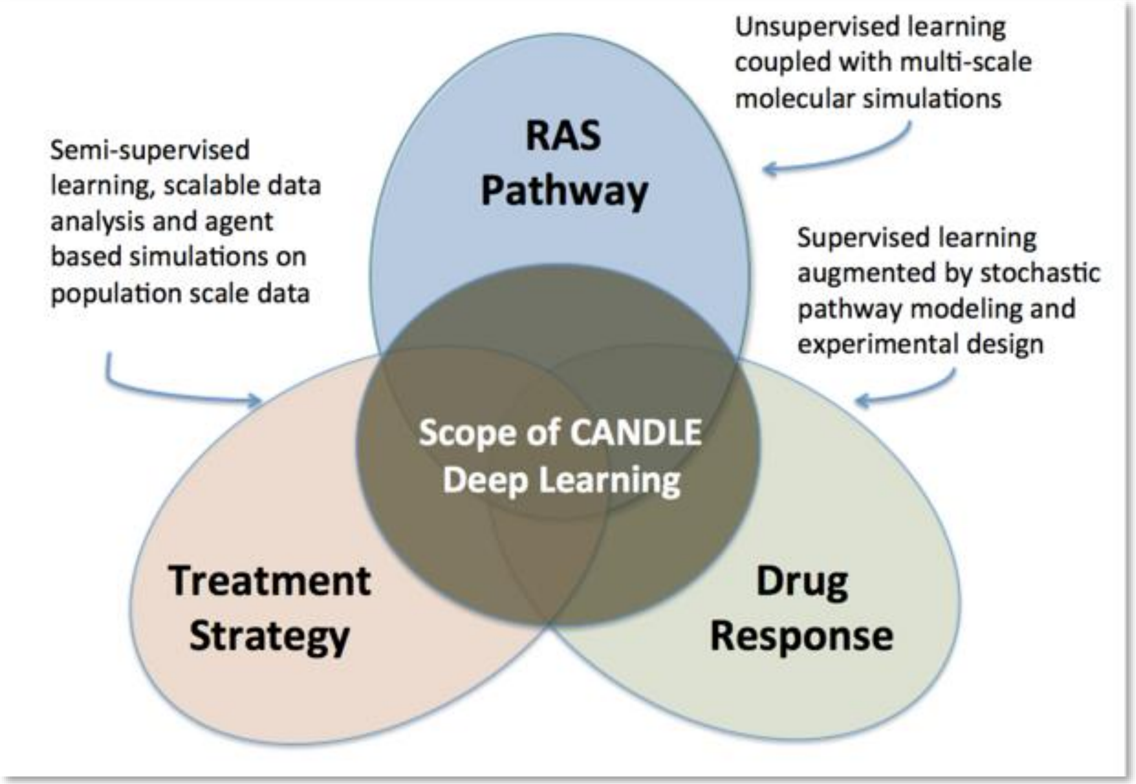
Gina Tourassi
Oak Ridge National Laboratory

Fred Streitz
Lawrence Livermore National Laboratory

Paul Dotson
Los Alamos National Laboratory

Special thanks to Tom Brettin, Harry Yoo, John Gounley, Jamal Mohd-Yushof, Emily Dietrich and the teams at ALCF, OLCF

ECP-CANDLE: CANcer Distributed Learning Environment



Unsupervised learning coupled with multi-scale molecular simulations

Semi-supervised learning, scalable data analysis and agent based simulations on population scale data

Supervised learning augmented by stochastic pathway modeling and experimental design

CANDLE Goals

Develop an exscale deep learning environment for cancer

Build on open source deep learning frameworks

Optimize for CORAL and exascale platforms

Support all three pilot project needs for deep learning

Collaborate with DOE computing centers, HPC vendors and ECP co-design and software technology projects

Cancer Distributed Learning Environment (CANDLE): An Exascale Computing Project Application Team

CANDLE Goals:

- **Build an Exascale computing environment for applying deep learning to problems in Cancer research supporting the three DOE/NCI Pilots**
- **Enable DOE's Leadership Computers to support scalable Deep Learning research** by building on top of industry leading deep learning frameworks (TensorFlow and PyTorch) and industry leading GPUs, NVIDIA (Summit), AMD (Frontier), and Intel (Aurora)
- **Provide open-source "ready-to-use" tools and libraries** in Python for large-scale needs of deep learning (e.g., HPO, ensembles, uncertainty quantification, workflows, validation, parallel training and inference, data pre/post processing)
- **Reduce the barriers for biomedical research community to utilize Exascale computing capabilities** and other HPC systems for deep learning

CANDLE Project: Findings

- CANDLE is an ECP project to enable the challenging deep learning problems in cancer research to be pursued on the supercomputers in the DOE. The scope of the CANDLE project includes identifying treatment strategy, drug response and RAS pathway for cancer related challenges, with a focus on a common software framework for these three areas. The goals of the project includes developing an exascale class open source deep learning environment for cancer and to collaborate with DOE computing centers, HPC vendors and ECP to co-design software technology projects.

- CANDLE has been delivering its milestones for the last six years and has performed well at the recent ECP annual review. The project has enabled fundamental science research and the team was involved in three Gordon Bell finalist nominations. This has demonstrated the ability to use large scale supercomputers as a viable scientific instrument to solve challenging and complex problems for a broader positive impact on humanity.

- The team has also participated in numerous outreach activities in the form of hackathons and tutorials to encourage wider community adoption of the tool sets developed. The code developed as part of this project is available via open source repository for evaluation. In addition, there is evidence of vendor adoption (NVIDIA) and the environment is provided as a container for exploring the possibility of the CANDLE.

- CANDLE has been installed at the NIH as part of its computing infrastructure, Biowulf, and is available for intermural NIH investigators and their collaborators. CANDLE is also part of the acceptance tests for the next generation of DOE Exascale systems, including Aurora and Frontier. CANDLE was also among the first large-scale machine learning codes in the Exascale Software Program.

CANDLE Project: Comments

- The team has been outstanding in their ability to pivot and make the necessary changes to support Covid related research during the pandemic, thereby demonstrating the potential of this tool chain beyond cancer.
- Enabling a supported and maintained CANDLE environment in a cloud platform will help scientists and engineers to continue evaluating the tool for a variety of models outside the supported hackathon and tutorial environments.
- The tool is mature having gone through multiple years of development effort. The team is encouraged to consider the sustainability of the CANDLE framework. We anticipate more models and domains using this tool chain and it is important to consider how to manage the core set of benchmarks and to remain agile whilst serving the wider cancer community needs.

CANDLE Project: Recommendations

Broader impact

On-board more grass-root level researchers in the cancer domain and continue lowering the barrier to entry for non-experts in computing

Democratize tools

Rapid evolution of computer hardware means medium size systems are very powerful science instruments - consider broadening accessibility by deploying on mid-clusters (as has already been done for Biowulf@NIH)

Evidence of impact

Highlight specific use cases of direct impact on cancer through the use of tool chain to further substantiate the value of this outstanding tool chain

Summary: DOE-NCI Projects

MOSSAIC Project:

The committee felt that this is an outstanding team effort which is laying the foundation for highly impactful information extraction efforts in a broad range of biomedical clinical, research and surveillance efforts.

ADMIRRAL Project:

The ADMIRRAL project is an important study investigating the dynamics of RAS-RAF interactions with a detailed and time-resolved approach. The project has successfully brought together an outstanding set of analytic capabilities, linked to high-performance computational modeling from the DOE team at LLNL, with the impressive breadth of experimental techniques from the experimental team at NCI's Frederick National Laboratory for Cancer Research.

IMPROVE Project:

This is a well-structured project that focuses on achievable deliverables that address the mission of both agencies. The project has completed an impressive set of actions and deliverables, and is now actively seeking to engage more with the community. The ultimate goal is to provide a generalizable framework that will enable the scalable comparison and improvement of existing deep learning models, regardless of the application science, underscoring the importance of making the protocols and software broadly available.

Summary: CANDLE ECP Project

The CANDLE project is a mature ECP project that has delivered on all its milestones over the last six years. With three finalist nominations for Gordon Bell awards, the project has clearly demonstrated the potential for large-scale supercomputers to perform leading-edge science. The CANDLE framework has also now been installed on the NCI Biowulf facility.