Report of ASCAC Subcommittee for DOE-NCI Collaborative Program

Tony Hey

June 2023

Charge Letter to ASCAC

- The 2016 Memorandum of Understanding for the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) 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

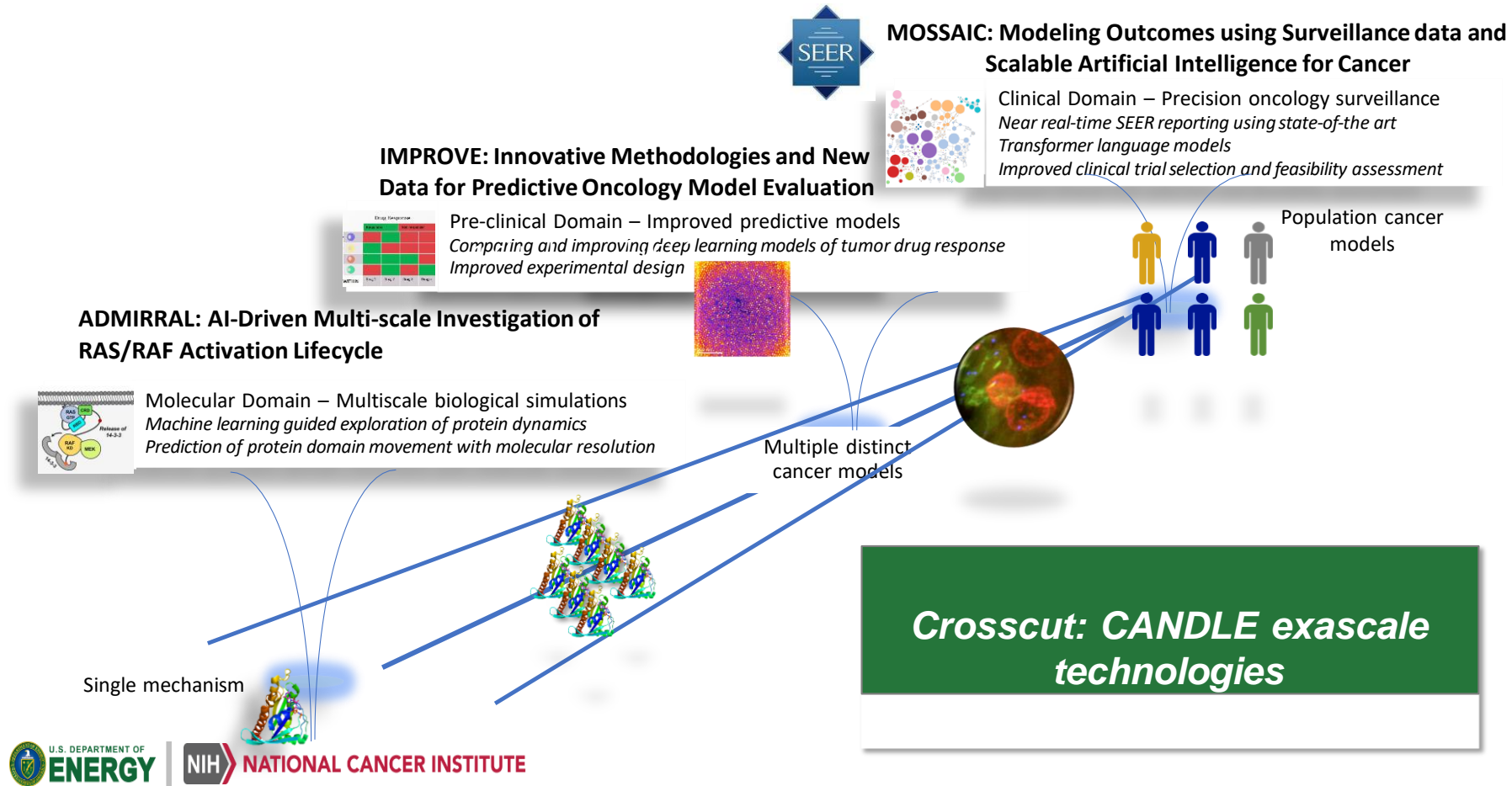|	NCI-DOE Collaborations ASCAC Subcommittee	NCI-DOE Collaborations Scientific & Technical Advisory Committees	NCI-DOE Collaborations Executive Committee
Number	1	1 per project	1
Member composition	Chair: Tony Hey. Members: 8-12 extramural scientists with expertise across collaboration areas (cancer, biology, advanced computing, data science, etc.)	4-6 scientists per committee with targeted, deep expertise relevant to the assigned project	NCI: Drs. Bertagnolli, Lowy, Singer, Barnholtz-Sloan, Goddard 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 – Amazon UK

- 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

Population cancer models

Multiple distinct cancer models

Single mechanism

Crosscut: CANDLE exascale technologies

SEER

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

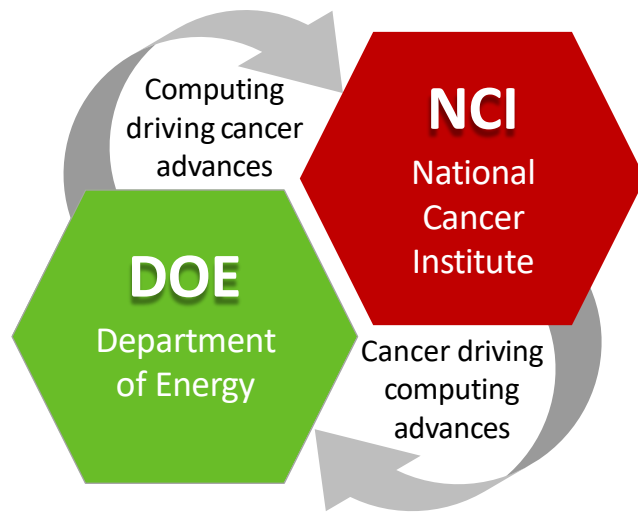# DOE-NCI Project Reviews in 2023

- Meeting at Frederick National Laboratory on June 6th and 7th
  - Hybrid meeting with 7 subcommittee members attending in person and 2 on Zoom

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

- The 3 DOE-NCI Projects
  - MOSSAIC – Joel Saltz and Caroline Chung
  - ADMIRRAL – Jay Bardhan and Amanda Randles
  - IMPROVE – Rick Arthur and Susan Gregurick

- Final review of cross-cutting ECP Project
  - CANDLE – Satheesh Maheswaran and Martin Berzins

Draft Project Reviews

June 6th and 7th

MOSSAIC Review

Presentation given by Betsy Hsu (NCI), Heidi Hanson (ORNL),
and Ola Adeyemi (FNL)

Lynne Penberthy (NCI) and Gina Tourassi (ORNL) in virtual attendance
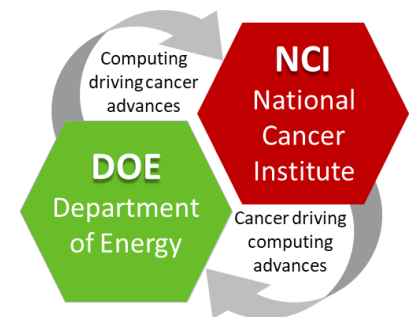
Findings:

- Extracts clinical information - cancer diagnosis, pathology, biomarkers, initial and follow-on treatments
- Automates cancer information abstraction for NCI SEER cancer registries and assesses SEER eligibility
- Ongoing development of recurrence and metastasis prediction model
- Extremely high positive predictive values along with uncertainty quantification
- Deployed in 16 SEER sites – roughly 31% of US population; software now deployed as default as part of any new DMS installation, regardless of SEER affiliation
- Validation – multiple VA sites; lead by University of Utah
  - Autoencoding performance increased from 17% to 23-27% of path reports with > 98% accuracy across all data elements
  - For remaining 75% of cases, highest probability elements offered to coders to improve operational efficiency. Emory SEER demonstrated use of API in SEER*DMS Pathology Screening Interface
- Rapid case ascertainment (RCA) studies near real-time data to identify patients rapidly for studies
- Open source pipelines – Framework for Exploring Scalable Computational Oncology FrESCO
- Research in federated learning with differential privacy at the level of models

Comments:

- Outstanding work in development of a demonstrably scalable NLP clinical data extraction system

- The currently deployed model is a Multi-Task Hierarchical Self Attention Network (HiSAN) – available open source. Ongoing development of a transformer-based foundation model

- Foundation model trained using self-supervised learning and fine tuned for additional tasks. This will allow easy generalization to other data sources and tasks

- API deployed – several compelling examples of API use to improve operational efficiency in GA SEER registry.

- Methods leveraged for use in near real-time reporting and rapid case ascertainment for research studies

- Collaboration with CDC to develop a privacy preserving API

- Leverages ORNL's HIPAA compliant HPC environment tools; recent successful run of MOSSAIC workflows on 2048 nodes of Frontier

Recommendations:

- The transition to self-supervised foundation models is an excellent approach and the team enthusiastically agrees with plans to leverage foundation models to broaden set of targeted clinical nodes and discrete data elements

- The subcommittee agrees with plans to apply approach to extraction of clinical data in many additional translational research settings. Efforts to evaluate and improve data quality are encouraged, starting with some pilot studies with initial collaborative sites

- The subcommittee encourages continued work towards adapting MOSSAIC for use in rapid case ascertainment for research studies and near real-time incidence reporting

- The subcommittee encourages continued effort to engage new partners for collaborative development and for deployment of MOSSAIC tools

- As the ability to easily target MOSSAIC methods to new data elements improves, it would be useful to know how the performance characteristics of the MOSSAIC pipelines compare with pipelines created by other clinical informatics groups. Collaborative comparative assessments of pipelines are encouraged

- **Summary:** the subcommittee felt that this is an outstanding team effort. Excellent progress has been made over the past year in core methods development, in generalization of methods to new application challenges, and in developing collaborations with new collaborators.
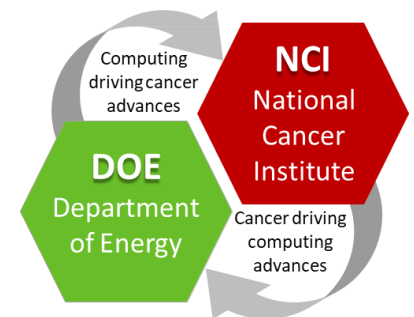
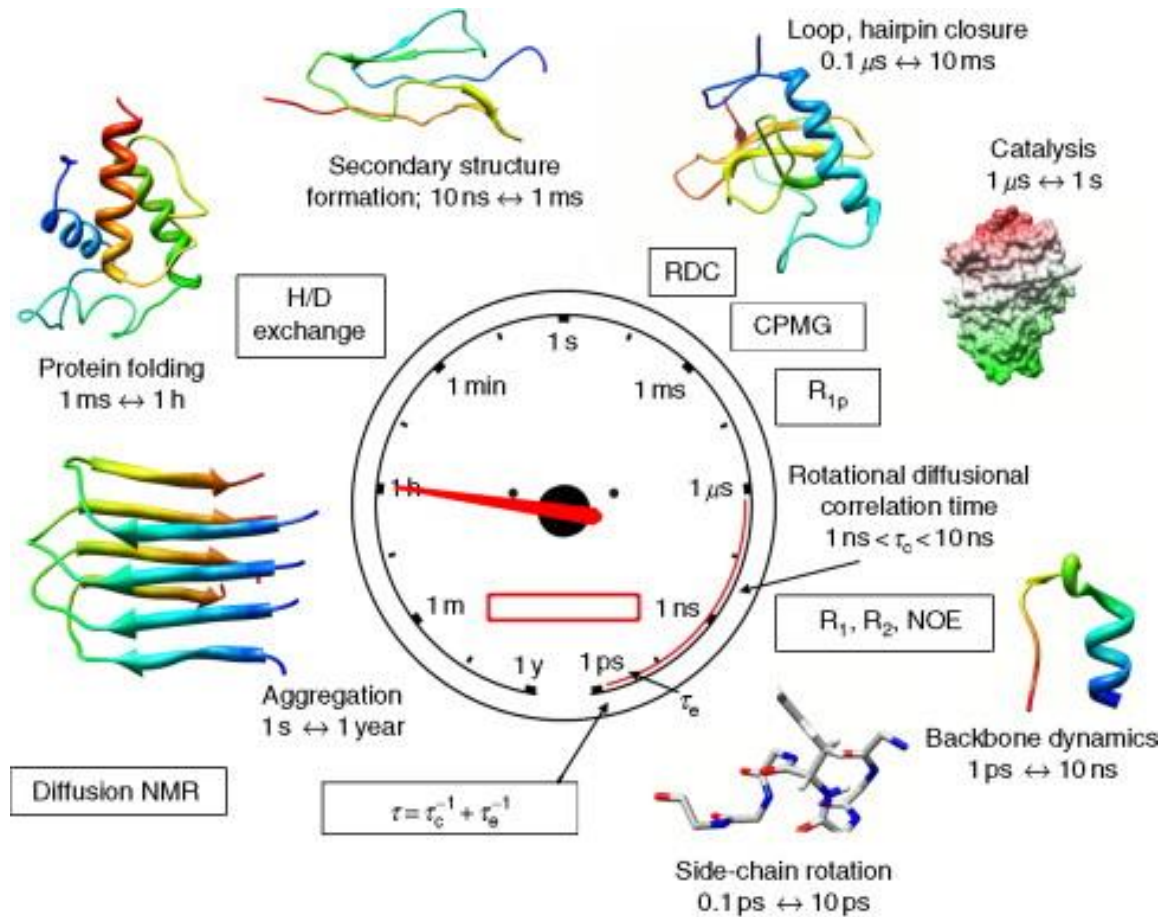# ADMIRRAL: AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle

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

- 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**
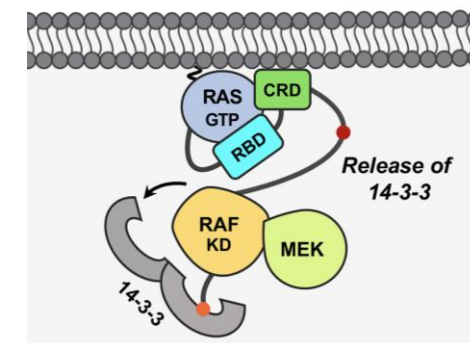
Overall Goal: investigate the longer length- and time-scale conformation changes that are at the heart of activation of the RAS-RAF complex.

Subcommittee Findings:

- Significant progress improving the LLNL ddcMD multi-physics particle dynamics code and incorporating the Martini-3 lipid potentials

- Multiscale modeling approach of MuMMI v.1 has been upgraded substantially with the ultra-coarse-graining (UCG) approach and hypothesis-generation engine

- Experimental validation to date is appropriate for the specific problem and the plans for future experiments are strong

ADMIRRAL Project

Presented milestone status update

ADMIRRAL	Status
Evolving latent space exploited for hypothesis generation	ML model built and tested to allow generation of possible structures
Coarsest scale is hypothesis generation engine	MuMMI refactored to enable queries to arbitrary generation engines
Create anisotropic lipid potentials to handle RAS/RAF local operations	2D lipid potentials for new macro model developed using CGMD simulations
Upgrade to Martini 3	ddcMD refactored to utilize Martini 3 potentials
Need to define structure of RAF and RAS/RAF complex	Initial state (inactive RAF in contact with RAS on membrane) defined; secondary state in process

ADMIRRAL Project
Relative to 2022 Subcommittee Recommendations:

Recommendation	Status
The in vitro analysis with purified proteins and membranes is important and highly developed. However, this does not explore two important biological complexities, whose relevance to function are not well understood.	The team is aware of considerations about heterogeneity, as exemplified by their comparison of DDFT and CG simulations for RAF-lipid interactions in membranes of differing composition. The team has made compelling progress towards identifying the transition between states, and after this is substantially complete, the team should return to this recommendation, even if only to explain the challenges of addressing it.
An important question that could have greater emphasis is how the models will be tested with experimental studies.	Key experimental work critical to justifying the created models has been performed (e.g. characterizing 7 isoforms of 14-3-3 and their affinity for the 2 BRAF phosphorylation site). Some elements of the coarse-grained modeling have been appropriately validated with finer-grain models. The plans for near-term experimental validation are strong and should provide meaningful insights.
The progress recommends a re-assessment of opportunities and priorities, in large part because of the incredible progress, but also due to the rate of progress in other areas, e.g. AlphaFold and the final deliverables of ECP.	This does not appear to have been pursued explicitly, however, the team has clearly considered how the methodologies and tools developed here apply to other use cases, not only in cancer biology (for instance, other players in the RAS/RAF pathway) but also in abiotic settings, such as materials science and engineering.
It would be valuable to consider the challenges and scope of effort required to harmonize the software stack with the ECP stack.	It is not clear if this has been pursued. However, the subcommittee understands that between the previous briefing and this one, the team lost two of its central computational scientists, which cost them significant time.
The project would benefit from engaging the broader research communities relevant to key technologies employed by this project.	The team includes an expert in statistical mechanics who is working to advance theory and practice connecting the multiscale approaches of ADMIRRAL to theoretical frameworks like Transition Path Sampling.

Comments:

- Performance portability is a concern – substantial efforts needed for porting to new machines making sustainability a potential issue, given both the expected diversification of accelerators, and evolutions of force fields (e.g. AI/ML derived force fields where inference is performed at each time step)

- For computational results obtained recently, experiments provide vital support to the meaningfulness of the computing work, and planned future experiments represent a wide variety of techniques that should offer a comprehensive view validating the computations and providing additional insights.

- Recent extensions to the MuMMI approach should substantially improve capabilities (Martini2->3 and UCG for hypothesis generation). With that said, there needs to be much more comprehensive experimental work to assess the generality of their developed modeling frameworks. This is true for the UCG and the lipid-protein anisotropy model.

- There are exciting possibilities for applications across biology (NCI, DOE, and others) and elsewhere that multiscale molecular-to-mesoscale modeling is needed

Recommendations:

- Follow through on the breadth of planned experiments

- Develop and implement a plan to engage relevant modeling communities (e.g. DOE BER/BES)

- Identify at least one additional molecular machine relevant to cancer for which the MuMMI v.2 would be transformative and perform the necessary campaign preparations for that system. Document the process and challenges encountered thoroughly to ensure the methodology's general applicability.

- Perform additional experimental studies to assess the strengths and limitations of the critical methodological advances (ultra-coarse-grained model, particularly with an implicit membrane; the macro-model of anisotropic lipid protein interactions; and the use of latent variables to support identifying transition paths). These studies will provide a documented, sound basis for defining clear guidelines for appropriate usage, which is essential for community engagement and sustainment.

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

Project Review

June 6th, 2023

Presentation by Rick Stevens (ANL), Ryan Weil (FNL), Neeraj Kumar (PNNL), Sarah Gosline (PNNL), and Nick Chia (Mayo)
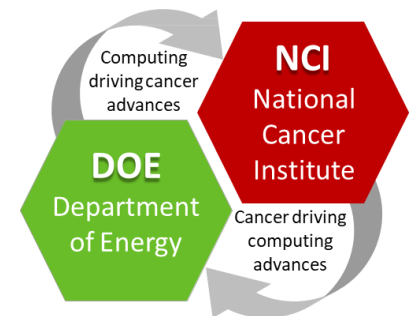
Overall Goal: A community-based, "automated" framework for model cross-comparisons

Subcommittee Findings:

- **Active** collaboration with Mayo Clinic, Texas Tech University, and Pacific Northwest National Laboratory
- Scientific Advisory Committee with representation from computer science, AI, and cancer communities
- Sixteen published AI models in the IMPROVE Framework, evaluating these pan-cancer and pan-drug models for single-drug response prediction
  - Requires model curation and hyperparameter optimization for each model to be cross-compared
  - Exploring optimization and benchmarking across differing computational configurations/infrastructures
- Developing a method to understand on how a model's prediction accuracy varies across different drug/tumor features.
  - **Surprising early finding**: the method to encode the drug structure in the AI model illustrates some models perform better on certain drug features. This could have ramifications in drug design by AI
- Curated and generated benchmark data for evaluating drug response prediction performance within and across datasets.
- (Jan 2023 alpha release): Codebase and documentation of IMPROVE on Github, data and curated models available via ftp

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

Relative to 2022 Subcommittee Recommendations:

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.	✓ Focusing on Colon Adenocarcinoma (with some related histological subtypes), designing seed study using PDOs & PDCs as cancer models for screening
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.	✓ Model curation and hyperparameter optimization workflows developed, benchmarked data developed ✓ Cross-study generalization analysis workflow developed ❑ Publication of methodology uncertain/not presented
Continue community engagement with researchers studying early tumor development, an opportunity for keen insight into the types of data that may strengthen predictive models.	✓ Expanded end user community – shown results from Mayo Clinic looking at breast, hepatobiliary, lymphoma, multi-myeloma, ovarian cancers ✓ Engaging Vendors for additional data ❑ Expand to new partners beyond close collaborators
Consider developing metrics to baseline and form a measure for assessment and comparison inclusive of factors such as productivity, completeness, correctness, and consistency.	✓ New CMP-CV comparison framework with cross-validation: types of errors, model characteristics providing better predictive accuracy, impact of features of datasets (e.g., # of atoms, bonds, etc.) and model robustness to noise in the data. ❑ Unaddressed? Productivity / usability of models

Comments:

- Early engagement across stakeholders via scientific advisory board and (private) **hackathons**, plans for public hackathons 4Q23, 1Q24.

- Good progress in structuring and characterizing of both models and data – more difficult to clearly determine progress against project goals.

- Consider: use case(s) for **patient digital twin** (selection of model to apply to an *individual* as opposed to corpus/population) e.g., tracking improvement vs. expectation/prediction as well as use case(s) targeting (or extracting) subpopulation/phenotype applicability for **precision medicine**.

- Consider: propose design of (and possibly pilot) a "**data market**" (incentives/subsidies) to stimulate data capture in regions of elevated uncertainty and/or conspicuous sparsity.

- Consider: Incentives for Cancer AI community to **contribute models before publication**, at least for benchmarking and optimization, if not for **distinction of novelty and impact**.
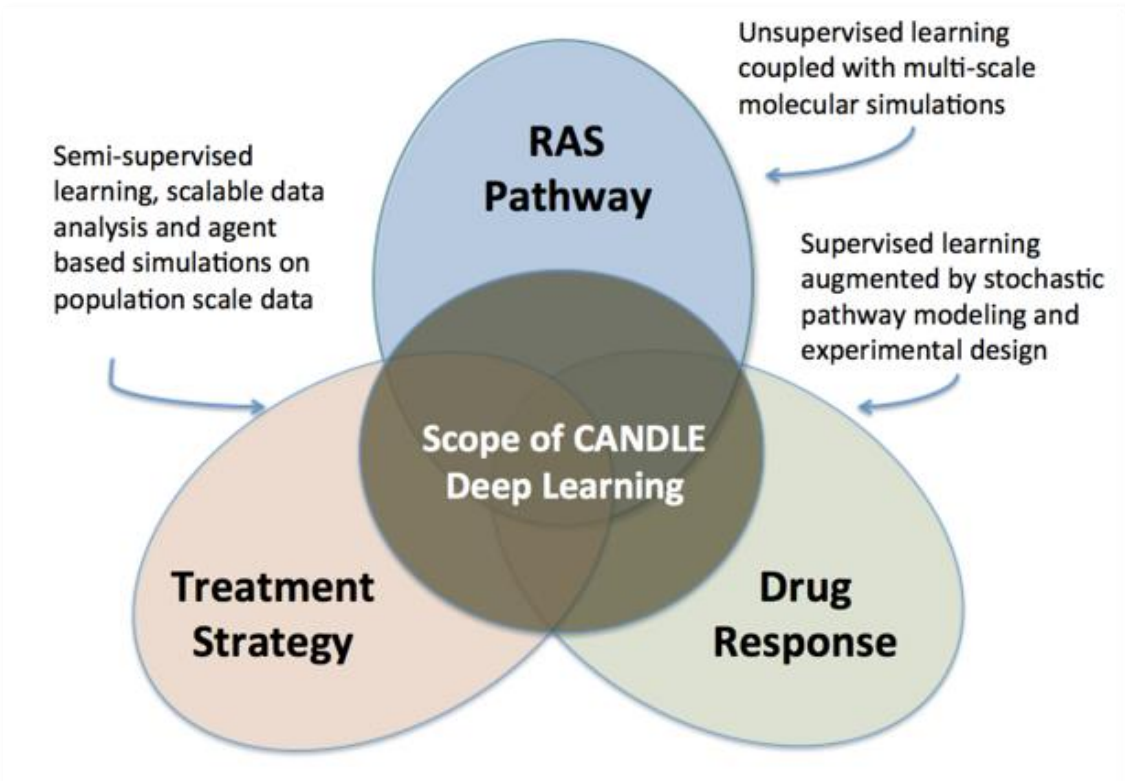
Recommendations:

- **Build 2-3 clear and compelling results stories based on outcomes** (even if notional/aspirational) and then work backwards to describe how the tools and workflows support.

- **Clarify data feature impact results** (molecular properties, etc.) – focus on 1-2 examples in depth, then "zoom out" to demonstrate wider number of features assessed, curious / important observations across these factors.

- Intuitive (to end-user) and consistent means of **formally characterizing regions of model competence** / performance scoring of models within target n-D region of competence/**domain of applicability**, assist in selecting potential to combine of models for regions of interest.

- Examine **formalisms** being developed in industry **for model characterization** (e.g., metadata schema in UMC4ES from NAFEMS-ASSESS) and interoperability (OMG SysML, FMI / DCP)

- Consider generation of **phantom (synthetic as-visualized/reported-in-practice) predictive results**, as would be reviewed by end user physicians, oncologists, pathologists, radiologists, etc.

- **Publications** of project results including methodology developed, cross-model comparisons and insights into / from datasets, and suggestions in employing IMPROVE in practice across potential communities.

ECP-CANDLE: CANcer Distributed Learning Environment

Unsupervised learning coupled with multi-scale molecular simulations

RAS Pathway

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

Supervised learning augmented by stochastic pathway modeling and experimental design

Scope of CANDLE Deep Learning

Treatment Strategy

Drug Response

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

CANDLE project

- Project funded as part of ECP and ends in December 2023

- Goal to deliver a viable Exascale-optimised software framework for Deep Learning applied to Cancer and other potential drug discovery scenarios.

- All milestones were delivered on time
  - Covid research delivered on top of agreed Candle milestones – this demonstrated the potential to transfer technology developed for cancer into other domains – prestigious awards were awarded for this work(?)
  - Used for about 30+ Cancer Deep learning Models

- Developed to leverage Exascale systems and has demonstrated success at scale on Frontier and Summit so far.

- The KPP benchmarking performance improvement of 50 was exceeded by 5X

- Explores the development of Large Language Models for science domains – agreed with program office to take on this extra work

CANDLE project

- Code openly <u>available</u> on GitHub
  - FTP site hosts all <u>public</u> data sets
- CANDLE helped bring cancer research communities together with hands-on workshops
- Tens of thousands of downloads but not clear as to usage
- Project team has a credible software sustainability plan
- Delivered against two broad paradigms - hyperparameter optimisation and ensemble learning
- Learning and tool chains adopted successfully by the IMPROVE project, specifically offering hyperparameter optimisation as a service (alpha release in March 2023)
- <u>Overall a significantly successful research and software project</u>

CANDLE: Stretch Goal (Rick Stevens)

- A stretch goal of the CANDLE project is to develop a DOE Transformer-based modeling framework that can run at scale

- In Fall of 2020 the CANDLE team joined forces with the ExaLearn ECP project to investigate LLMs

- LLM foundation models were run on Frontier and Aurora prototypes before the ChatGPT revolution

- These LLM transformers were trained on scientific datasets

- The AI4SES town meetings in 2022 identified opportunities to use foundation models for science

- Developing international consortium to develop 1 Trillion parameter model using NVIDIA Megatron and Microsoft DeepSpeed codebase for training