Report of ASCAC Subcommittee for DOE-NCI Collaborative Program

Tony Hey

September 2023

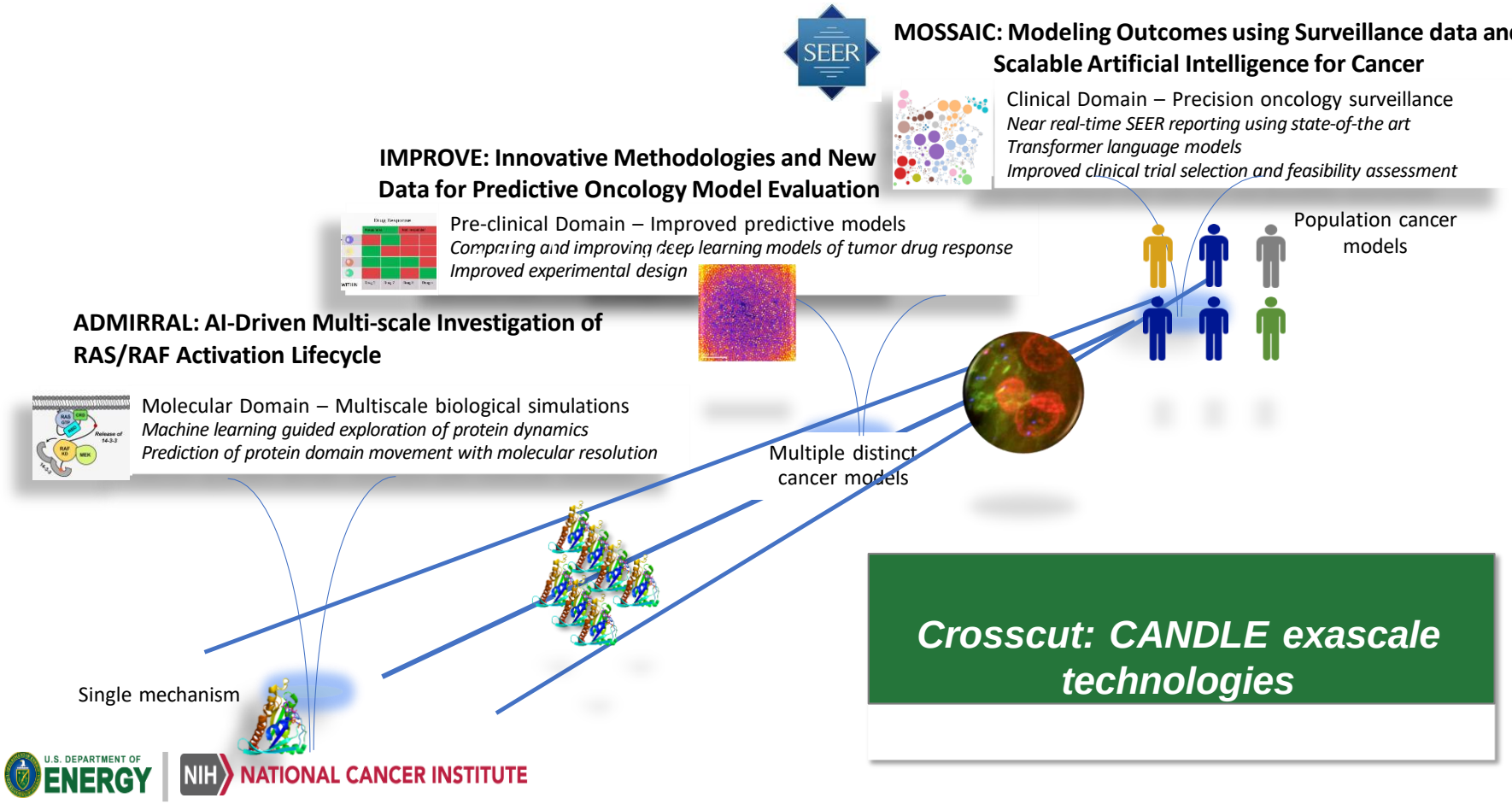
Charge Letter to ASCAC

- 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.

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



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

Executive Summary

- The MOSSAIC project most mature and impactful of the three projects. Its technology is already being used in 16 SEER sites and the VA registry. The subcommittee was pleased to see exploration of the use of ‘foundation models’ to SEER data.
- The ADMIRRAL project is ambitious exploration of Multiscale Machine-learned Modeling Infrastructure (MuMMI) computational model of the RAS/RAF complex. Project has made significant progress this year in introducing more powerful AI-enhanced modeling capabilities. Experimental validation of the model predictions is impressive.
- The IMPROVE project is only in its second year but has already made significant progress towards creating a community-based framework for cross-comparison of AI models used to validate cance drug response models. The codebase and documentation are available on Github with data and curated models.

Presentation on ECP CANDLE project

- Goal to deliver a viable Exascale-optimised software framework for Deep Learning applied to Cancer and other potential drug discovery scenarios.
- Provided Deep Learning benchmarks for all three DOE-NCI projects
- All milestones delivered on time
  - Developed to leverage Exascale systems and has demonstrated success at scale on Frontier and Summit so far.
  - KPP benchmarking performance improvement of 50 was exceeded by 5X
  - Code and data on GitHub
  - Used for about 30+ Cancer Deep learning Models
- Additional achievements
  - Covid research delivered on top of agreed Candle milestones – this demonstrated the potential to transfer technology developed for cancer into other domains
  - GenSLM (Genome-Scale Language Models) revealed SARS-CoV-2 Evolutionary Dynamics: paper awarded Gordon Bell special prize at SC22

CANDLE: Stretch Goal (Rick Stevens)

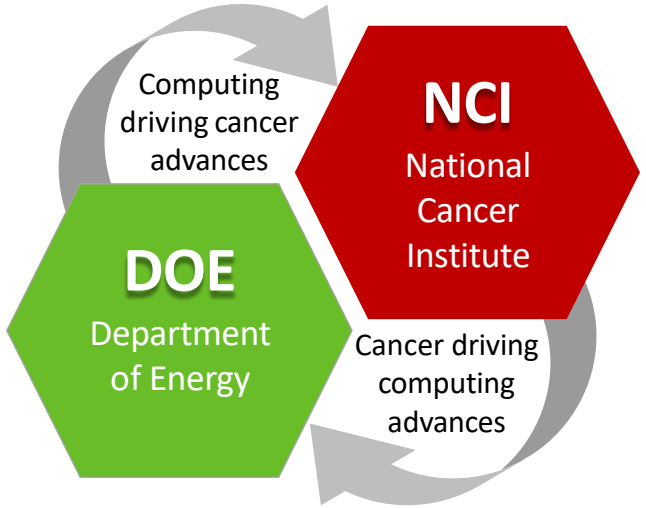
- Stretch goal of the CANDLE project is to develop a DOE Transformer-based modeling framework that can run at scale
- In Fall of 2020 the CANDLE team joined forces with the ExaLearn ECP project to investigate LLMs
- LLM foundation models trained on scientific datasets were run on Frontier and Aurora prototypes before the ChatGPT revolution
- The AI4SES town meetings in 2022 identified opportunities to use foundation models for science
- Developing international TPC consortium to develop 1 Trillion parameter model using NVIDIA Megatron and Microsoft DeepSpeed codebase for training

Detailed Project Reviews

June 6th and 7th

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

DOE-NCI partnership to advance exascale development through cancer research



June 7, 2023

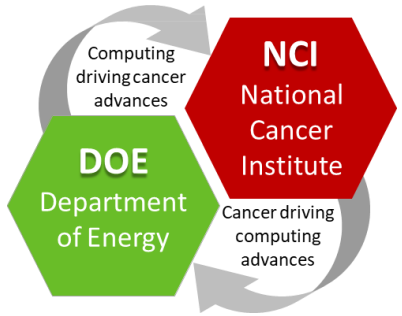
Presentation to:
ASCAC DOE-NCI Subcommittee

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 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:

- Automates cancer information abstraction for NCI SEER cancer registries and assesses SEER eligibility
- Extracts clinical information - cancer diagnosis, pathology, biomarkers, initial and follow-on treatments
- 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; led 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 as FrESCO. 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:

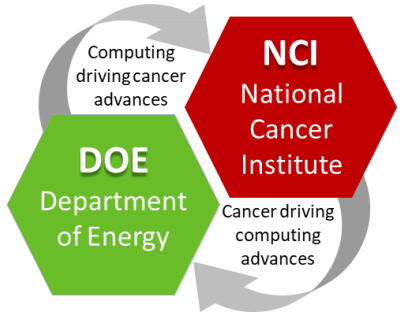
- The transition to self-supervised foundation models is an excellent approach and the team enthusiastically agrees with plans to leverage foundation models to broaden set of targeted clinical nodes and discrete data elements
- The subcommittee agrees with plans to apply approach to extraction of clinical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

Project Review

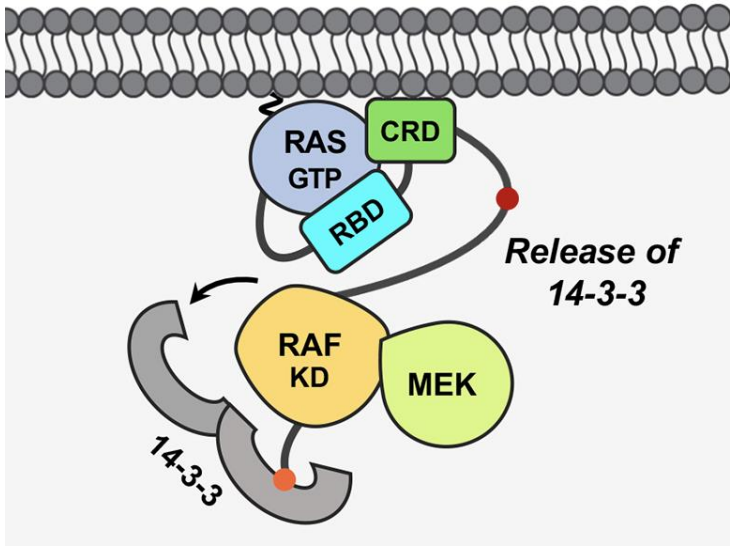
June 6-7, 2023

Presentation given by Dwight Nissley (FNL)
and Fred Streitz (LLNL)



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

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

Findings:

The most significant advances include:

- 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

The team has delivered on all the key milestones:

- exploiting the latent space to generate possible new structures;
- re-factoring the MuMMI code to enable the use of arbitrary hypothesis generation engines;
- developing and validating new anisotropic potentials for lipids for more accurate RAS/RAF interactions with them;
- re-factoring ddcMD to utilize Martini 3 (see later); and
- defining an initial state of RAF/RAS on a membrane.

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:

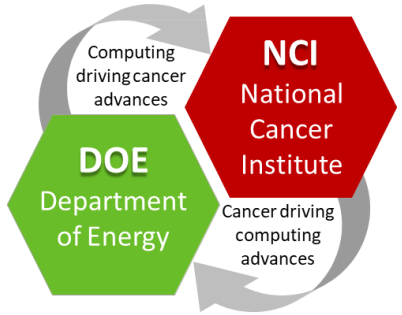
- Follow through on the breadth of planned experiments
- Develop and implement a plan to engage relevant modeling communities (e.g. DOE BER/BES)
- Identify at least one additional molecular machine relevant to cancer for which the MuMMI v.2 would be transformative and perform the necessary campaign preparations 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)



IMPROVE Overall Goal: A community-based, "automated" framework for model cross-comparisons

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
- 16 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

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**.

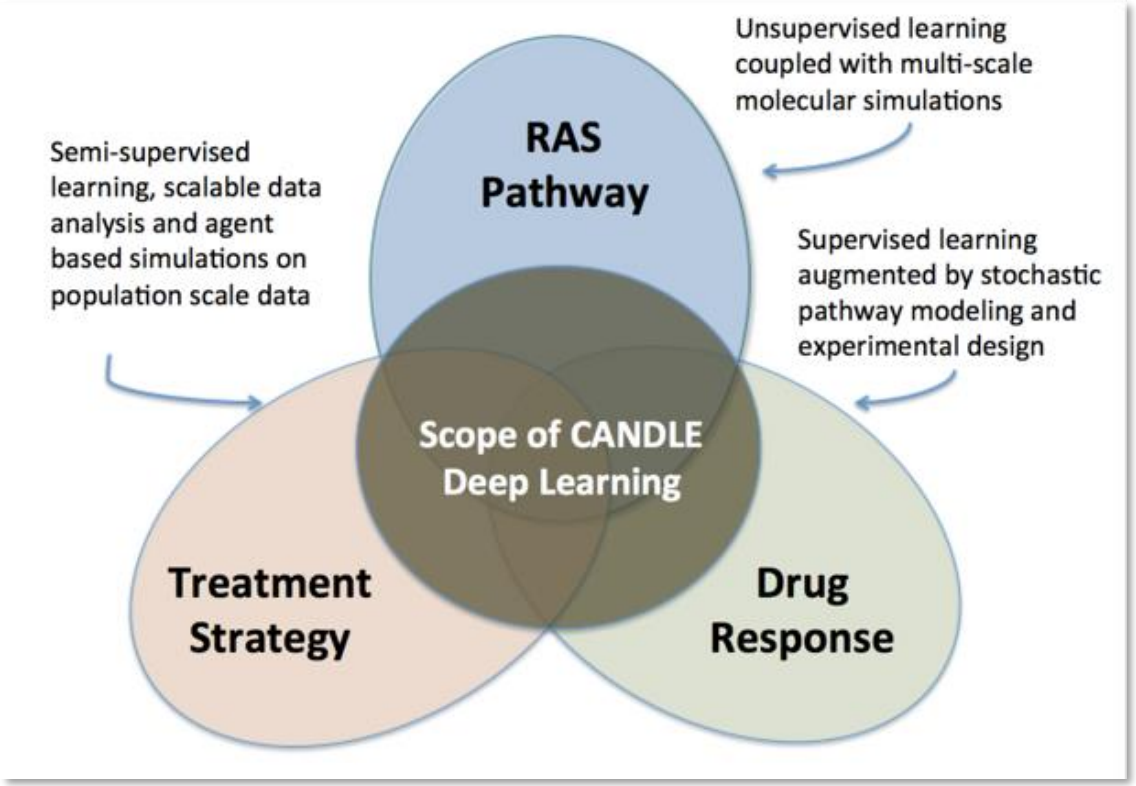
Recommendations:

- **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.

CANDLE ECP project
Informal review

Presentation by Rick Stevens (ANL), John Gounley (ORNL),
and Tom Brettin (ANL)

ECP-CANDLE: CANcer Distributed Learning Environment



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

NATIONAL CANCER INSTITUTE

NVIDIA IBM CRAY intel AMD

ECP EXASCALE COMPUTING PROJECT

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 award for this work(GenSLM at SC22)
  - 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
- Stretch goal to explore development of Large Language Models for science domains – agreed with program office to take on this extra work

CANDLE project

- Code openly available on GitHub
  - FTP site hosts all public 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)
- Overall a significantly successful research and software project

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