U.S. Department of Energy
Advanced Scientific Computing Advisory Committee
Subcommittee Report on the DOE-NCI Collaboration

Report to the Office of Science

September 2023

Professor Tony Hey

Subcommittee Chair

Contents

Executive Summary

The subcommittee was impressed by the achievements of all three of the Department of Energy-National Cancer Institute (DOE-NCI) projects and felt strongly that this was clearly a successful and interesting collaborative program.

- The **M**odeling **O**utcomes using **S**urveillance data and **S**calable **AI** for **C**ancer (MOSSAIC) project is the most mature, and the subcommittee was particularly pleased to see the project's exploration of the use of 'foundation models' leveraging deep learning applied to the Surveillance, Epidemiology, and End Results (SEER) data.
- The **AI-D**riven **M**ulti-scale **I**nvestigation of **R**AS/**R**AF **A**ctivation **L**ifecycle (ADMIRRAL) project is perhaps the most ambitious of the three projects. This year, despite some staffing problems, the project has made significant progress towards producing more powerful AI-enhanced simulation and modeling capabilities for the RAS/RAF complex. The experimental validation of the Multiscale Machine-learned Modeling Infrastructure (MuMMI) model is also very impressive and clearly important.
- The **I**nnovative **M**ethodologies and New Data for **PR**edictive **O**ncology model **EV**aluation (IMPROVE) project is at a much earlier stage than the other two projects but has already made significant progress towards creating a community-based framework for cross-comparison of deep learning models used to validate cancer drug response models. The codebase and documentation are now available on Github with data and curated models.

Introduction

The Department of Energy-National Cancer Institute (DOE-NCI) Collaboration, started in June 2016, is an interagency strategic partnership enabled by the National Strategic Computing Initiative and Precision Medicine Initiative. DOE contributes advanced computing and artificial intelligence (AI) expertise along with high performance computing (HPC) resources. NCI contributes driving cancer research questions along with subject matter expertise and sources of data. Overall, five DOE national laboratories are involved in this effort (Argonne National Laboratory, ANL; Brookhaven National Laboratory, BNL; Lawrence Livermore National Laboratory, LLNL; Los Alamos National Laboratory, LANL; and Oak Ridge National Laboratory, ORNL), Frederick National Laboratory for Cancer Research (FNL), and multiple NCI divisions and centers.

This major effort stems from the Joint Design for Advanced Computing Solutions for Cancer (JDACS4C) program, which is aimed at advancing cancer research through application of large-scale computing while providing design input to DOE for future large-scale computing systems and AI technologies through their application to complex, highly heterogeneous datasets. JDACS4C originally encompassed three pilots focused on problems at the molecular, cellular, and population levels. In 2021, with the signing of a new memorandum of understanding (MOU) between DOE Office of Science, DOE National Nuclear Security Administration, and NCI, the pilots transitioned to more flexible and sustainable projects (**I**nnovative **M**ethodologies and New Data for **PR**edictive **O**ncology model **EV**aluation, IMPROVE; **AI-D**riven **M**ulti-scale **I**nvestigation of

RAS/**R**AF **A**ctivation **L**ifecycle, ADMIRRAL; and **M**odeling **O**utcomes using **S**urveillance data and **S**calable **AI** for **C**ancer, MOSSAIC) with updated goals. Together, these projects are intended to bring a new class of advanced computing capabilities to bear on cancer research through the development of a shared information technology ecosystem and novel analytical algorithms.

The first challenge, the IMPROVE project, is to develop a framework for comparing and evaluating deep learning predictive models for drug response that can be used to optimize preclinical drug screening and drive precision medicine-based treatments for cancer patients. The second challenge, the ADMIRRAL project, is to understand the molecular basis of key protein interactions in the RAS-RAF-MEK-ERK signaling pathway that regulates the proliferation, differentiation and survival of cells. The RAS protein is mutated (constitutively activated) in about 30% of all human cancers. The third challenge, the MOSSAIC project, is to automate the analysis and extraction of information from millions of cancer patient records to ultimately help determine optimal cancer treatment strategies across a range of patient lifestyles, environmental exposures, cancer types, and health care systems.

Although these challenges are at different scales and have specific scientific teams collaborating on the data acquisition, data analysis, model formulation, and scientific runs of simulations, they also share several common threads. In addition, the Exascale Computing Project's (ECP's) CANDLE project — **CAN**cer **D**istributed **L**earning **E**nvironment — is deploying a single scalable deep neural network (DNN) code, scaled to exascale resources, which can be demonstrated on the three cancer use cases.

Under the 2021 MOU, reviews of these efforts shifted from the Frederick National Laboratory Advisory Committee to the Advanced Scientific Computing Advisory Committee (ASCAC). As a result, a subcommittee working group was formed, and the following is a summary of our second-year review. The charge letter to the chair of ASCAC and a list of subcommittee members are attached as Appendices 1 and 2, respectively.

IMPROVE Project Review

PI: Rick Stevens, ANL

Co-PIs: Jeff Hildesheim, NCI and Ryan Weil, FNL

*Progress on the IMPROVE project — Innovative Methodologies and New Data for PRedictive Oncology Model **EV**aluation — was presented by Rick Stevens (ANL), Ryan Weil (FNL), Neeraj Kumar (Pacific Northwest National Laboratory, PNNL), Sarah Gosline (PNNL), and Nick Chia (Mayo Clinic).*

The Subcommittee lead reviewers were Rick Arthur and Susan Gregurick.

Findings

The project is tasked to create a community-based (and accepted) "automated" framework for cross-comparison of predictive models. While the literature in the AI-cancer modeling field has grown considerably over the past several years, at the inception of the project the community

did not have a common set of well-documented and well-characterized approaches to constructing, training, and validating cancer drug response models. IMPROVE aims to address this shortcoming and is structured with two aims:

1. to develop a scalable, generalizable framework to compare deep learning models from the cancer drug modeling community, and to identify model attributes that contribute to a models' prediction performance; and
2. to develop 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.

While the methodologies should ultimately be effective more generally, currently the context of the study focuses on oncology within the DOE-NCI collaborations. The effort has convened a Scientific Advisory Committee with representation from the computer science, AI, and cancer communities, and evolved active collaborations including the Mayo Clinic, Texas Tech University, and PNNL, who presented results to the ASCAC-appointed review committee.

While at a much earlier stage than the other reviewed projects, the IMPROVE project shared early progress and several interesting and important accomplishments. The project has curated and generated benchmark data for evaluating drug response prediction performance within and across datasets. The codebase and documentation are available on Github, with data and curated models available via file transfer protocol (ftp). Adoption was promoted through private hackathons, with public hackathons planned for 4Q23 and 1Q24.

At the time of the review, the IMPROVE framework had published and evaluated 16 pan-cancer and pan-drug models, applying AI for single-drug response prediction. These studies required model curation and hyperparameter optimization for each model to be cross-compared, as well as exploration of optimization and benchmarking across differing computational configurations/ infrastructures. The team is working toward developing a method to understand how a model's prediction accuracy varies across different drug/ tumor features. One surprising early finding is that the method to encode the drug structure in the AI model illustrates that some models perform better on certain drug features. This could have ramifications in drug design by AI. Additional accomplishments include application programming interfaces (APIs) for generalized analysis across studies and datasets, including prediction performance, learning curve analysis, usefulness, and influence of intrinsic noise on prediction.

Comments

This section reviews how recommendations from the 2022 review have been addressed by the project:

- The focus on colon adenocarcinoma (with some related histological subtypes) — designing a seed study using patient-derived organoids (PDOs) and patient-derived tumor cell cultures (PDCs) as cancer models for screening — addressed the recommendation to focus and show a specific and rigorous application (then subsequently generalize). While

showing progress in clearly structuring and characterizing both models and data, directly assessing progress against project goals remained difficult.

- For a methodical generalizability strategy (cancer-to-cancer, to non-cancer, to non-biomedical), the project developed workflows for model curation and hyperparameter optimization as well as benchmarked datasets. However, this methodology was not presented in detail and should be published/ presented in the future.
- The collaboration has expanded, including engagements with vendors for additional data to strengthen predictive capabilities of models, as well as with the framework's user community — including results presented by the Mayo Clinic on breast, hepatobiliary, lymphoma, multi-myeloma, and ovarian cancers. Further expansion to new partners remains encouraged.
- To consider a broader set of evaluation metrics for models, the project shared a new Comparison-Cross Validation (CMP-CV) framework: 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. These are important core metrics, but others with value as incentives for adoption remain unaddressed, including usability of the models (impacting productivity, consistency, and utility).

Recommendations

The 2023 subcommittee review team makes some further recommendations to the IMPROVE team:

1. Build two or three clear and compelling results stories based on outcomes (even if notional/aspirational) and then work backwards to describe how the tools and workflows support the analysis. For example, use case(s) for patient digital twins (selection of model to apply to an individual as opposed to corpus/ population) and tracking improvement vs. expectation/prediction, as well as use case(s) targeting (or extracting) subpopulation/phenotype applicability for precision medicine. Another means to better frame an outcome story could be generation of phantom (synthetic as-visualized/ reported-in-practice) predictive results, as would be reviewed by end-user physicians, oncologists, pathologists, radiologists, etc.
2. Clarify data feature impact results (molecular properties, etc.) focusing on one or two examples in depth, then "zoom out" to demonstrate the wider number of features assessed, and curious/ important observations across those factors. In regions of elevated uncertainty and/ or conspicuous data sparsity, consider designing (and possibly piloting) a "data market" (incentives/ subsidies) to stimulate value-targeted data capture.
3. Develop and deliver an 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 to assist users in selecting potential models (and combinations of models) for regions of interest.
4. Toward more comprehensive model description and broader application, consider leveraging formalisms being developed in industry for model characterization (e.g., metadata schema in UMC4ES from NAFEMS-ASSESS) and interoperability (OMG SysML, FMI / DCP).

5. Ultimately, publish project results including methodologies developed, cross-model comparisons and insights into/ from datasets, and suggestions in employing IMPROVE in practice across potential communities. Engage with publishers across the Cancer (/AI) community to suggest incentives for contributing models for benchmarking and optimization, (if not for distinction of novelty) prior to publication.

Summary

The IMPROVE project, while relatively new, has developed workflows for AI model curation and for hyperparameter optimization as well as a comparative framework as a means to benchmark applications in developing and applying pan-drug AI models for single-drug response predictions. The team consists of AI experts and experts in oncology and drug design who work in a collaborative manner. The team has delivered benchmarked datasets, a codebase and documentation (on Github) and has tested their methodology with 16 pan-drug AI-models with interesting initial findings. The team connects with the community through a Scientific Advisory Committee and through active collaborations including the Mayo Clinic, Texas Tech University, and PNNL. The team has also conducted hackathons.

The ASCAC subcommittee found that the IMPROVE team had been responsive to the first-year review and has made improvements to focus on specific applications in cancer — here on colon adenocarcinoma — and in developing new collaborations. To have impact in the AI-cancer community, the ASCAC review committee recommends that the team publish its methodology and findings, leverage model characterization formalisms from industry, develop impactful use-case stories, and clarify initial data-feature results.

ADMIRRAL Project Review

Co-PIs: Fred Streitz, LLNL, and Dwight Nissley, FNL

Progress on the ADMIRRAL project — AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle — was presented by Dwight Nissley (FNL) and Fred Streitz (LLNL).

The Subcommittee lead reviewers were Jay Bardhan and Amanda Randles.

Findings:

The ADMIRRAL project continues to make impressive progress towards an AI-enhanced simulation and modeling capability and applying it to understand the dynamics of RAS-RAF interactions in cancer. Since the last progress report issued by this subcommittee, the team has achieved its key milestones on both the computational and experimental fronts and taken steps to address the recommendations in the previous report. While the team remains robust, it is important to note that attrition of key early- and mid-career technical experts has contributed to slowed progress.

The most significant advances include:

- Experimental validation to date is appropriate for the specific problem, and the plans for future experiments are strong;
- Significant progress has been made in improving the LLNL ddcMD multi-physics particle dynamics code and incorporating the Martini-3 lipid potentials; and
- The multiscale modeling approach of MuMMI (Multiscale Machine-learned Modeling Infrastructure) v1.0 has been upgraded substantially with the ultra-coarse-graining (UCG) approach and hypothesis-generation engine.

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.

The subcommittee's primary recommendations from the 2022 review have been diligently addressed by the team. First, the team has been advancing the modeling capabilities to address membrane heterogeneity (mainly through DDFT and CG simulations). Second, the team undertook essential validation efforts in response to the subcommittee's encouragement to connect more with experimental measurements and high-fidelity simulations. These efforts included experimental characterization of seven isoforms of the 14-3-3 protein and their respective affinities for the two BRAF phosphorylation sites. Additionally, the team has meticulously assessed specific elements of the coarse-grained modeling methods through comparisons with fine-grain models. Their additional plans for near-term experimental validation are strong and are anticipated to yield meaningful insights.

Comments:

- Performance portability seems to be a concern for long-term sustainment. The team reports substantial efforts are needed for porting to new machines, e.g., for optimizing on Frontier. Given the expected diversification of accelerators and ongoing evolutions of simulation force fields (e.g., AI/ ML-derived force fields where inference is performed at each time step), a strategic sustainment plan would secure the investment's value.
- 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, additional, comprehensive experimental work is needed to assess the generality of their developed modeling frameworks. This need is also 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.

- The project is premised on the hypothesis that understanding the conformational change in RAS-RAF on activation is critical to identifying and developing therapeutics. While it is evident that this work will provide insights into conformational changes, it remains to be seen if, when, or how its relevance to therapeutics will be assessed. The project, therefore, seems on a stronger, more impactful footing to focus on improving the fundamental understanding of large-scale molecular machines in cancer.

Recommendations:

1. The team should follow through on the breadth of planned experiments. Furthermore, wherever possible, we encourage the team to conduct 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.
2. We encourage the team to continue its efforts to respond to the 2022 recommendation to re-assess opportunities and priorities for the collaboration. Due to the substantial progress in building the infrastructure, the most impactful use of resources may be identifying new applications or engaging the community to adopt the AI-enhanced framework for discovery (see below). It may be valuable for the team, the project's Science Advisory Board, and the subcommittee to convene and discuss the Science Advisory Board's perspectives. Given the complexity of the modeling tools and workflow, it would substantially help make these investments of long-term value if the team identified 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. By documenting the process (including engagement with domain experts) and challenges encountered, the methodology's general applicability will be made more apparent to the scientific community, and undoubtedly improvements will also be part of the process.
3. Concerning engaging the community, other projects reviewed by the subcommittee have substantial engagement activities which should serve as a useful model. It is encouraging that the team now includes an expert in statistical mechanics who can connect the multiscale approach to rigorous theory and existing frameworks such as Transition Path Sampling, and further such contacts and points of leverage should be seized.

Summary:

The ADMIRRAL team continues to make impressive progress, enabling the investigation of the long time-scale and long length-scale conformation changes at the heart of the activation of the RAS-RAF complex. The new MuMMI framework surpasses the earlier model, enabling machine learning-guided exploration of protein motions in the complex milieu of heterogeneous cell membranes. These developments significantly move scientific capabilities forward towards a traditionally intractable biological challenge of predicting protein domain movement with molecular resolution and understanding such movement's effects on cellular decision making.

MOSSAIC Project Review

Co-PIs: Gina Tourassi, ORNL, and Lynne Penberthy, NCI

*Progress on the MOSSAIC project — **M**odeling **O**utcomes using **S**urveillance data and **S**calable **AI** for **C**ancer — was presented by Betsy Hsu (NCI) and Heidi Hanson (ORNL).*

The Subcommittee lead reviewers were Caroline Chung and Joel Saltz.

Findings

MOSSAIC is an ambitious project that targets extraction of detailed and nuanced clinical information concerning cancer diagnosis, pathology, biomarkers, initial and follow-on treatments, metastases, and recurrence. The immediate motivation for this work had been to automate cancer information abstraction for NCI SEER cancer registries in order to enable more consistent, efficient, and timely data abstraction. Specifically, the project is now also targeting "real-time" cancer reporting compared to the current approximately two-year lag of cancer reporting to SEER. Production implementation has been carried out in 16 SEER sites; roughly 31% of the U.S, population. Deployment has also been carried out in the Vertans Affairs (VA) registry. The VA trained the network created by MOSSAIC using VA data obtained from multiple VA sites. Cancer attributes include site, subsite, histology, and grade. The classifications encompass considerable nuance, with 70 site categories, 324 subsite categories, and 626 histology categories. The focus was initially to develop and validate methods that have extremely high positive predictive values, with uncertainty quantification used to identify a subset of reports that could be classified with very high accuracy. Specifically, for auto-extraction from pathology reports, the initial approach has been refined over the past year, resulting in improved performance of the percentage of pathology reports coded with over 98% accuracy across all data elements and increasing from 17% up to 23-27%.

The above described fully automated approach uses uncertainty quantification to identify a subset of reports that can be classified with very high accuracy. This allows the methods to be used to automate the coding that would otherwise be carried out by humans. The previously developed API has been leveraged to create an interface that allows users to select from high probability options; this is used to improve the operational efficiency by focusing manual coders on cases where the uncertainty estimation process is unable to select a single set of codes.

The project has also targeted the crucial issue of determining whether a case is reportable to SEER and is targeting rapid case ascertainment (RCA) studies using near real-time data to identify patients for studies. The investigators report a value of 0.997 for the reportability determination task. The rapid case ascertainment has the promise of supporting patient identification for research studies, clinical trials matching, near real-time incidence monitoring, and understanding the impact of new diagnostics and treatments to outcomes.

This group has been extremely proactive in the testing and adoption of natural language processing (NLP) deep learning methods, and continues to carry out extensive work on both benchmarking and model development. The currently deployed model makes use of a Multi-Task Hierarchical Self Attention Network (HiSAN). This network and associated software has been

publicly distributed as FrESCO — A framework for Exploring Scalable Computational Oncology. This is a modular deep-learning NLP library for extracting pathology information from clinical text documents. The FrESCO model was trained and employed on VA cancer registry data though a multi-disciplinary collaboration that includes multiple VA sites and the University of Utah. Additional improvements were reported including interesting work on incorporating code hierarchy via attention.

The project is now developing foundation models (Path-BigBird) that target extraction of new oncology elements. The foundation models are trained in a self-supervised manner followed by tuning for multiple downstream tasks. The downstream tasks being targeted include biomarkers and recurrence along with the currently predicted site, subsite, histology, behavior, and laterality categories. Foundation models will allow greater flexibility in choosing additional downstream tasks. These models will be used for additional tasks such as analysis of radiology reports, clinical notes, treatment details, and extraction of data that characterize recurrent disease and presentation of metastases. Initial performance benchmarking is demonstrating small to moderate performance improvements in existing tasks and increases in accuracy for tumor subsite and behavior.

The investigators have also launched studies to assess for algorithm bias and noise. For algorithm bias, the initial use case was around race bias in HER2 status and breast cancer. Specifically, there were noted differences in F1 scores by race with black women noted to have a two-fold increased risk of triple negative breast cancer and even when presenting with non-metastatic disease, were found to be more likely to die of their disease compared to their white counterparts.

MOSSAIC has been able to leverage the ORNL Health Insurance Portability and Accountability Act- (HIPAA-) compliant HPC environment tools. It uses Oak Ridge Leadership Computing Facility- (OLCF-) developed data transfer protocols and the CITADEL framework for secure Summit access and scalable analyses. The project has made extensive use of Summit and is being migrated to Frontier, with migration to three Frontier development environments — Tulip, Spock and Crusher — and a recent run on the 2,048 Frontier nodes. The CITADEL framework is being migrated to Frontier.

To expand model access to the broader community without compromising confidentiality, MOSSAIC is carrying out two parallel efforts. The first effort makes use of a secure archive — CITADEL — used to house a model and to carry out training and predictions. CITADEL is linked with sites through secure mechanisms for passing information back and forth to carry out model training and prediction. This approach requires cooperative agreements between participants and institutions and has been deployed across multiple sites. The other approach involves a differential privacy federated training approach which uses algorithms that trade off accuracy for privacy. The initial method studied by the group — differentially private stochastic gradient descent (DPSGD) — was found not to exhibit satisfactory performance. The group is actively exploring a different class of method — exponential mechanism with normalizing flows — which they anticipate will have better privacy/ accuracy tradeoffs.

Comments

- The project is conducting outstanding work in development of a demonstrably scalable NLP clinical data extraction system that also utilizes uncertainty quantification to identify subsets of reports that can be classified with very high accuracy and help focus manual abstractors to more difficult cases.
- The currently deployed model is a Multi-Task HiSAN that is available in open source.
- The foundation models under development are being trained using self-supervised learning and fine-tuned for additional tasks. This will greatly facilitate generalization to other data sources and tasks, with targets for data extraction including radiology reports, and oncology clinical nodes to extract mCODE data elements, for example.
- API deployed — several compelling examples of use of the API to improve operational efficiency in the Georgia (GA) SEER registry were presented.
- Methods are being leveraged for use in real-time reporting, rapid case ascertainment for research studies, and near real-time incidence reporting.
- There is collaboration with CDC to develop a privacy-preserving API.
- The project leverages ORNL HIPAA-compliant HPC environment tools; there was a recent successful run of MOSSAIC workflows on the 2,048 nodes of Frontier.
- The investigators aim to integrate environmental data into their analytic pipeline; this will use residential history data found in SEER sources and build environmentally linked models of cancer risk and cancer outcomes

Recommendations

1. The transition to self-supervised foundation models is an excellent approach, and the team enthusiastically agrees with plans to leverage foundation models to broaden the set of targeted classes of nodes and discrete data elements. Even with foundation models, integrating the current approach of leveraging quantification is encouraged.
2. The subcommittee encourages continued work towards adapting MOSSAIC for use in rapid case ascertainment for research studies and near real-time incidence reporting.
3. Beyond text-based data, the subcommittee encourages the MOSSAIC group to explore extending the foundation model infrastructure to whole-slide pathology and radiology image analysis tasks. In these areas, the metadata of the data generation and pre-processing would be interesting to note as uncertainty quantification is carried out.
4. The subcommittee enthusiastically endorses use of the CITADEL secure architecture to support broader community access to trained models. This will likely become even more compelling as the group transitions to creation and use of foundation model approaches.
5. The subcommittee encourages continued effort to engage new partners for collaborative development and for deployment of MOSSAIC tools.
6. The subcommittee lauds the principle investigators' (PIs') performance benchmarking efforts along with their work in assessment and amelioration of algorithm bias. The committee supports continued efforts in the detection and amelioration of algorithm bias and noise.

7. 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 cross comparison of automated, semi-automated and manual approaches may be informative for a publication out to the community and in prioritizing future directions.
8. The subcommittee endorses plans to enable SEER exposome analyses by making use of residential history information in SEER data sources to incorporate residential data in cancer risk and outcome analyses.

Summary

The subcommittee felt that the MOSSAIC project is an outstanding team effort. Excellent progress has been made over the past year in both core methods development, in generalization of methods to new application challenges, and in developing collaborations with new collaborators.

CANDLE Project Presentation

PI: Rick Stevens, ANL

This presentation was not subject to a formal review since this was an ECP project funded by DOE Office of Advanced Scientific Computing Research (ASCR).

*The CANDLE project —**CAN**cer **D**istributed **L**earning **E**nvironment — was presented by Rick Stevens (ANL), Tom Brettin (ANL), and John Gounley (ORNL). Brettin and Gounley presented slides on CANDLE project performance and Stevens presented slides on a CANDLE stretch goal to develop a DOE transformer-based modeling framework that can run at scale.*

The Subcommittee lead reviewers were Martin Berzins and Satheesh Maheswaran.

CANDLE Project Goals

- Develop an exascale deep learning environment for cancer
- Build on open-source deep learning frameworks
- Optimize for the Collaboration of Oak Ridge, Argonne, and Livermore (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

Comments

- The CANDLE project is funded as part of the ECP and ends in December 2023. At the recent ECP Annual Review, the ECP review team found that the project had delivered all its milestones on time and, in addition, CANDLE has greatly exceeded its baseline Key Performance Parameters (KPPs) on exascale systems by more than 100x.

- One goal for CANDLE was to deliver a viable exascale-optimized software framework for deep learning applied to cancer and other potential drug discovery scenarios. The CANDLE framework is currently used for more than 30 cancer deep learning models. The framework was developed to leverage exascale systems and has demonstrated success at scale on Frontier and Summit so far.
- The project has also provided significant support for task-parallel workflows for the IMPROVE and MOSSAIC DOE-NCI projects. CANDLE also provides a python library, benchmarks, runtime software, and a framework that allows the leveraging of frameworks such as Tensorflow, Keras, Horovod, and PyTorch.
- The CANDLE code and datasets are all openly available on GitHub; an ftp site hosts all public datasets.
    - There have been tens of thousands of downloads but the project is not clear as to usage statistics;
    - The CANDLE project team has developed a credible software sustainability plan;
    - CANDLE delivered software to support two broad paradigms:
        - hyperparameter optimization;
        - ensemble learning;
    - Deep learning and tool chains have been successfully adopted by the IMPROVE project, which is now offering hyperparameter optimization as a service (alpha release in March 2023);
- The CANDLE project also helped bring cancer research communities together with hands-on workshops;
- As a stretch goal, the CANDLE team agreed with ASCR to take on the exploration of developing large language models (LLMs) for science domains. In the Fall of 2020, the CANDLE team joined forces with the ExaLearn ECP project to develop a transformer-based modeling framework capable of running at scale. LLM foundation models were run on Frontier and on Aurora prototypes before the ChatGPT revolution.
- COVID-19 research has also been delivered in addition to the agreed CANDLE milestones. This has demonstrated the potential to transfer technology developed for cancer into other domains. Scientists from ANL and a team of collaborators won the 2022 Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research for their new method of quickly identifying how a virus evolves. Their work in training large language models to discover variants of SARS-CoV-2 has implications to biology beyond COVID.

Summary

CANDLE is a mature and significantly successful research and software project that has been of great benefit to the three DOE-NCI projects. The work on the new stretch goal on LLMs is particularly impressive.

Conclusions

IMPROVE

- The IMPROVE project, while relatively new, has developed workflows for AI model curation and for hyperparameter optimization, and a comparative framework as a means to benchmark applications in developing and applying pan-drug AI models for single-drug response predictions. The team consists of AI experts and experts in oncology and drug design, who work in a collaborative manner. The team has delivered benchmarked datasets, a codebase, and documentation (on Github) and has tested their methodology with 16 pan-drug AI-models with interesting initial findings. The team connects with the community through a Scientific Advisory Committee and through active collaborations including the Mayo Clinic, Texas Tech University, and PNNL. The team has also conducted hackathons.
- The ASCAC review committee found that the IMPROVE team was responsive to the first-year review and has made improvements to focus on specific applications in cancer — here on colon adenocarcinoma — and in developing new collaborations. To have impact in the AI-cancer community the ASCAC review committee recommends that the team publish its methodology and findings, leverage model characterization formalisms from industry, develop impactful use-case stories, and clarify initial data-feature results.

ADMIRRAL

- The ADMIRRAL team continues to make impressive progress, enabling the investigation of the long time-scale and long length-scale conformation changes at the heart of the activation of the RAS-RAF complex. The new MuMMI framework surpasses the earlier model, enabling machine learning-guided exploration of protein motions in the complex milieu of heterogeneous cell membranes.
- These developments significantly move scientific capabilities forward towards a traditionally intractable biological challenge of predicting protein domain movement with molecular resolution and understanding such movement's effects on cellular decision making.

MOSSAIC

- The subcommittee felt that the MOSSAIC project is an outstanding team effort. Excellent progress has been made over the past year in both core methods development, in generalization of methods to new application challenges, and in developing collaborations with new collaborators.

CANDLE

- CANDLE is a mature and significantly successful research and software project that has been of great benefit to the three DOE-NCI projects. The work on the new stretch goal on LLMs is particularly impressive.

Appendix 1: Charge Letter

Department of Energy
Office of Science
Washington, DC 20585

July 27, 2021

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Senior Vice President for Academic Affairs
Professor of Computer Science and Electrical & Computer Engineering
The University of Utah
201 Presidents Circle, Room 205
Salt Lake City, Utah 84112-9007

Dear Professor Reed:

Thank you for your work as Committee Chair on the Advanced Scientific Computing Advisory Committee (ASCAC) and for the ongoing Committee of Visitors review. The ASCAC recommendations will help us to improve the management of this important program.

The Office of Science and the National Nuclear Security Administration began collaborating with the National Cancer Institute (NCI) in 2016 to dramatically accelerate progress in cancer research through the application of High Performance Computing and Artificial Intelligence. This collaboration, through the Memorandum of Understanding for the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C), has recently been renewed for an additional five-year period. ASCAC generally provides advice and recommendations concerning the Advanced Scientific Computing program in response only to charges from the Director of the Office of Science. However, the Committee's objectives and scope of activities include periodic reviews of elements of the Advanced Scientific Computing Research Program and recommendations based thereon. Considering this, I request the ASCAC form a working group to review the activities under this collaboration and to provide advice to the Office of Science regarding new opportunities that might contribute significantly to these efforts; to identify any major challenges that are preventing the efforts from delivering on their potential; and to provide recommendations for how the Office of Science might address these challenges.

I request the ASCAC establish this working group for the life of the collaboration or until this charge is withdrawn. The working group should report findings through the ASCAC in an open public meeting or videoconference meeting of the ASCAC, as needed, but at least 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.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

J. Stephen Binkley
Acting Director
Office of Science

Appendix 2: NCI-DOE Subcommittee members

Tony Hey (ASCAC member and Chair)
Chief Data Scientist
Rutherford Appleton Laboratory
Science and Technology Facilities Council
Harwell Campus, Didcot OX11 0QX, UK
tony.hey@stfc.ac.uk

Rick Arthur (ASCAC member)
Senior Director/Senior Principal Engineer
Advanced Computational Methods Research
GE Research Center
1 Research Circle
Niskayuna, New York 12309
arthurr@ge.com

Vadim Backman

Sachs Family Professor of Biomedical Engineering and Medicine
McCormick School of Engineering and Applied Sciences
Northwestern University
Technological Institute
2145 Sheridan Road
Evanston, IL 60208
v-backman@northwestern.edu

Jay Bardhan
Computational Scientist
Environmental Molecular Sciences Division
Pacific Northwest National Laboratory
P.O. Box 999
Richland, WA 99352
jaydeep.bardhan@pnnl.gov

Martin Berzins (ASCAC member)
SCI Institute
Department of Computer Science
72 South Central Campus Drive, Room 3750
University of Utah
Salt Lake City, Utah 84112
mb@sci.utah.edu

Caroline Chung
Associate Professor
Department of Radiation Oncology
The University of Texas MD Anderson Cancer Center
Houston, TX 77030
CChung3@mdanderson.org

Susan Gregurick (ASCAC member)
Director, Division of Biomedical Technology, Bioinformatics, and Computational Biology
National Institute of General Medical Sciences
National Institutes of Health
45 Center Drive MSC 6200
Bethesda, Maryland 20892
susan.gregurick@nih.gov

Bill Gropp

Director, National Center for Supercomputing Applications
Thomas M. Siebel Chair Professor in Computer Science
Computer Science Department
University of Illinois Urbana-Champaign
1205 W. Clark St., MC-257
Urbana, IL 61801
wgropp@illinois.edu

Amie Hwang
Assistant Professor of Clinical Population and Public Health Sciences
Keck School of Medicine of USC
2001 N. Soto St.
Health Sciences Campus
Los Angeles, CA 90089
amiehwan@usc.edu

Satheesh Maheswaran
Senior Specialist Solutions Architect
Amazon Web Services (AWS)
Amazon UK
drmahes@amazon.co.uk

Gordon Mills
Professor of Cell, Developmental and Cancer Biology, School of Medicine
Director of Precision Oncology, Knight Cancer Institute, OHSU
Wayne and Julie Drinkward Endowed Chair in Precision Oncology
Oregon Health and Science University
Portland, OR 97201
millsg@ohsu.edu

Amanda Randles
Alfred Winborne and Victoria Stover Mordecai Assistant Professor of Biomedical Sciences
Biomedical Engineering
Wilkinson Building
534 Research Drive
Durham, NC 27708
amanda.randles@duke.edu

Joel Saltz
Cherith Professor and Founding Chair, Department of Biomedical Informatics
Associate Director, Stony Brook Cancer Center
Stony Brook, NY 11794
Joel.Saltz@stonybrookmedicine.edu