OVERVIEW OF COLLABORATION BETWEEN THE DOE, NNSA, AND NCI

Emily J. Greenspan, PhD, National Cancer Institute

Advanced Scientific Computing Advisory Committee

July 29, 2021

NIH NATIONAL CANCER INSTITUTE

Topics

- *Collaboration background*
- *Governance and oversight*
- *Project overviews*
  - *Pilot 2: RAS Biology on Membranes*
  - *Pilot 3: Precision Cancer Surveillance*
  - *Innovative Methodologies and New Data for Predictive Oncology Model Evaluation (IMPROVE)*
  - *ATOM*
- *Developed capabilities*
- *New opportunities*

Collaboration Background

Brief History

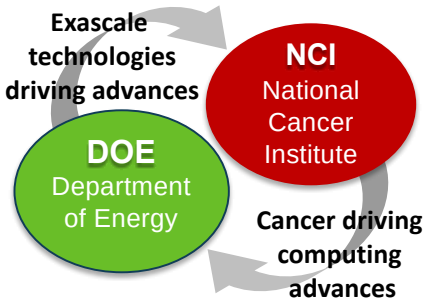
- **2014:** NSCI and PMI enabled potential for a joint High Performance Computing (HPC) focused collaboration between NCI and DOE
- **Winter – Spring of 2015:** started identifying a potential set of pilots
- **September 2015:** Presented to Frederick National Laboratory Advisory Committee (FNLAC; NCI FACA committee)
- **December 2015:** Presented to NCI Director Lowy and DOE Secretary Moniz
- **June 2016:** finalized original 5 year MOU between the two agencies
- **November 2019:** Pilot 2 wins best paper award at SC19
- **June 2021:** finalized new 5 year MOU between the two agencies (to be automatically renewed every 5 years)

JDACS4C Pilots: Pioneering New Computational Capabilities

Joint Design of Advanced Computing Solutions for Cancer (JDACS4C)

2016-present

NCI-DOE partnership to advance HPC through cancer research



Initiatives Supported NSCI and PMI

NIH NATIONAL CANCER INSTITUTE

Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY EST. 1943

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

Pilot 3: Population Information Integration, Analysis, and Modeling for Precision Cancer Surveillance

Clinical Domain – Precision oncology surveillance
Expanded SEER database information capture
Modeling patient health trajectories

PDMR NCI Patient-Derived Models Repository
An NCI Precision Oncology Initiative℠ Resource

Pilot 1: Predictive Modeling for Pre-Clinical Screening

Pre-clinical Domain – Improved predictive models
Computational/hybrid predictive models of drug response
Improved experimental design

Population cancer models

Pilot 2: RAS Biology on Membranes

Molecular Domain – Multiscale biological models
Models for RAS-RAS complex interactions
Insight into RAS related cancers

The RAS Initiative

Multiple distinct cancer models

Single mechanism

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

Crosscuts: CANDLE exascale technologies, Uncertainty Quantification (UQ)

Governance and Oversight

Original Governance and Oversight

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

	FNLAC NCI-DOE Collaboration Working Group	NCI-DOE Collaboration Governance Review Committee	FNLAC NCI-DOE Collaboration Task Force
Member composition	~15 extramural NCI and DOE funded scientists with expertise in computational and biological domain areas	~15 senior federal leadership and program leads from NCI and DOE	~12 extramural and intramural scientists with expertise in computational and cancer domain areas
Meeting Frequency	2-3 times per year	2-3 times per year	*Ad hoc* from May – Oct 2020
Charge/role	- Technical evaluation of JDACS4C pilot projects - Exploring additional areas for mutually beneficial NCI-DOE collaborations	- Guidance on strategy, policy, program management, budget - Communication with federal government and executive branch	- Review and assess merits of individual projects - Evaluate whether NCI-DOE Collaboration should continue and become a sustainable and stable partnership - Recommend future directions for NCI-DOE Collaboration

Overall Task Force Recommendations

- NCI-DOE Collaboration is uniquely suited to address certain critical challenges in cancer research and should continue
- The current pilots are really large, full-scale projects, and should be evaluated as such
- Future projects should be developed and reviewed by a more structured and rigorous approach
- Increase engagement with NCI extramural community
- Establish project-specific advisory groups

Project-Specific Task Force Recommendations

- **Pilot 1 should be concluded:** Insufficient available and pertinent data, insufficient integration with NCI investigators doing predictive modeling
- **Pilot 2 should be continued:** Greater focus on refining the coarse-grain models based on data from the atomic-level simulations and on experimental validation
- **Pilot 3 should be continued:** Greater focus on implementation and multi-institutional deployment of the developed APIs, and expansion beyond SEER

New MOU Governance and Oversight

	U.S. DEPARTMENT OF ENERGY | Office of Science	**NCI-DOE Collaborations Scientific & Technical Advisory Committees**	**NCI-DOE Collaborations Executive Committee**
Number	1	1 per project	1
Member composition	TBD	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	TBD	by project leads in consultation with Exec Committee	by agency leadership
Meeting Frequency	TBD	Quarterly or as needed	3 times per year
Charge/role	TBD	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

Project Overviews

Pilot 2: RAS Biology on Membranes

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

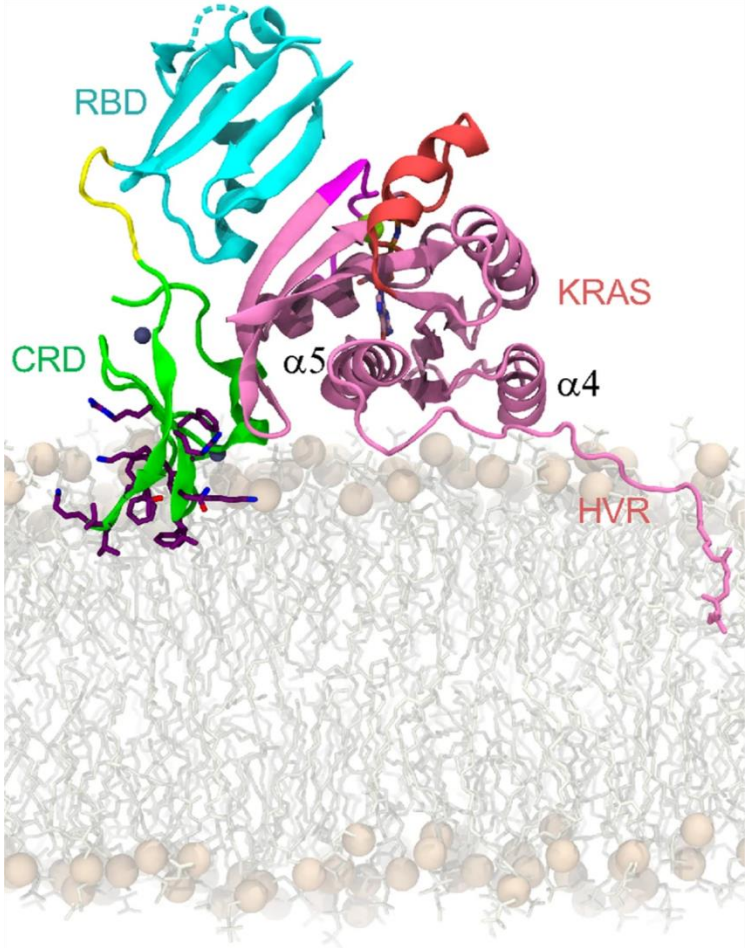


Fig. 7c from Tran, T.H., Chan, A.H., Young, L.C. *et al.* KRAS interaction with RAF1 RAS-binding domain and cysteine-rich domain provides insights into RAS-mediated RAF activation. *Nat Commun* 12, 1176 (2021). https://doi.org/10.1038/s41467-021-21422-x

Overall goal: deepen understanding of RAS biology and identify new drugs for RAS-related cancers through the integrated development and use of new simulations, predictive models and next-generation experimental data.

Aims:

- Understand how RAS and extended RAS complexes are activated
- Provide insight into RAS intracellular signaling
- Propose potential new therapies targeting RAS

Applies multi-scale molecular dynamics simulations of RAS-RAF interactions on realistic, lipid-bilayer membranes

Current capabilities (examples):

- Feature reduction of molecular dynamics simulation output using an autoencoder (P2B1)
- Simulation of curved bilayer membrane surfaces (MemSurfer)
- Determination of important samples in a constantly changing dataset (Dynim)

Pilot 3: Population Information Integration, Analysis, and Modeling (1)

Co-PIs: Gina Tourassi (ORNL), Lynne Penberthy (NCI)

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

Data:

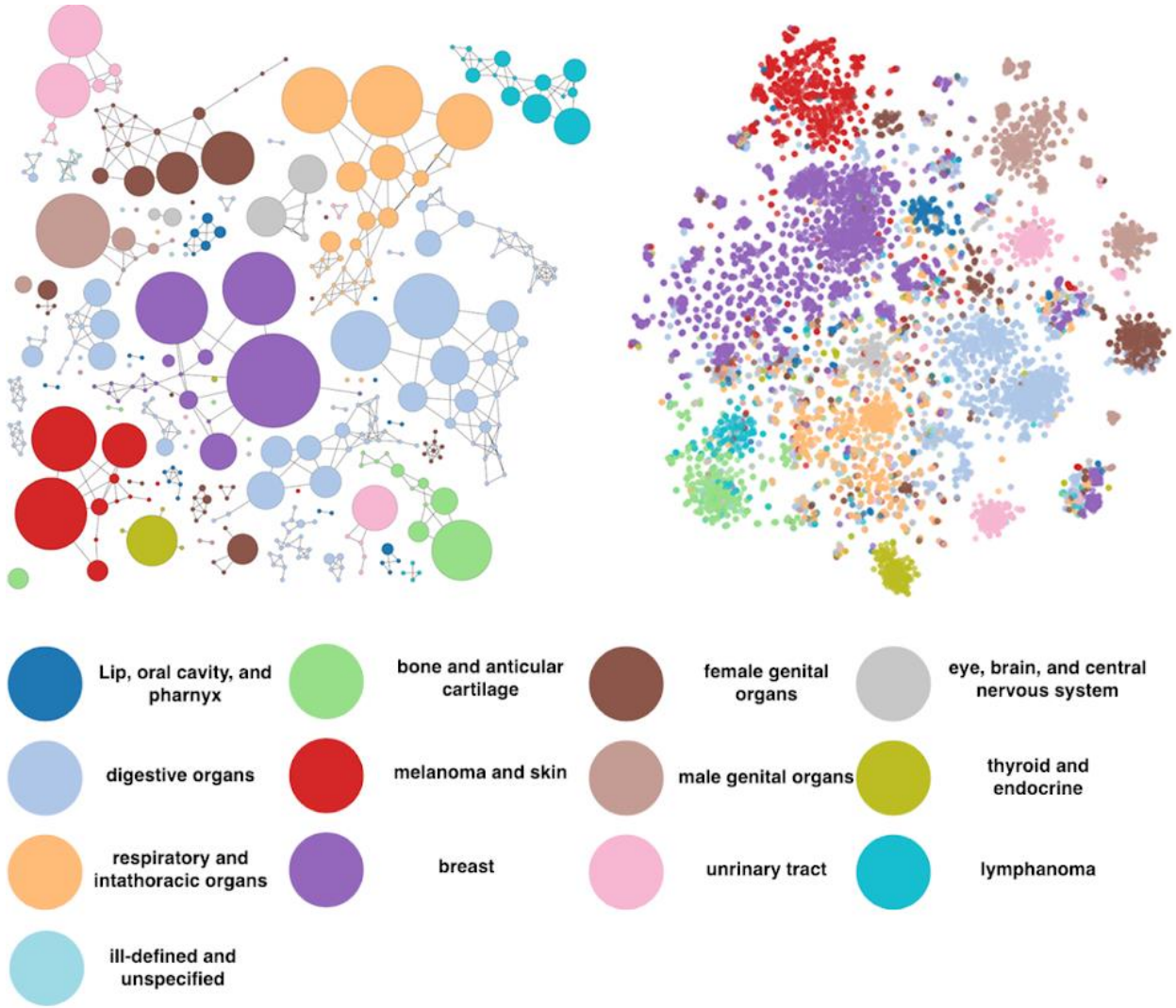
- 7 SEER cancer registries
- 1M unique cancer patients
- 3.5M cancer pathology reports
- Annotations for key data elements such as site, histology, behavior, etc.

Methods:

- Report-based and patient-based AI solution for information abstraction, reportability, recurrence
  - Convolutional neural networks
  - Self-attention networks
  - Large-scale Transformer language models
- Uncertainty quantification, abstention, active learning, knowledge distillation, etc.

Deployment:

- 15% workload reduction for Georgia cancer registry
- Test deployments in other SEER registries in 2021

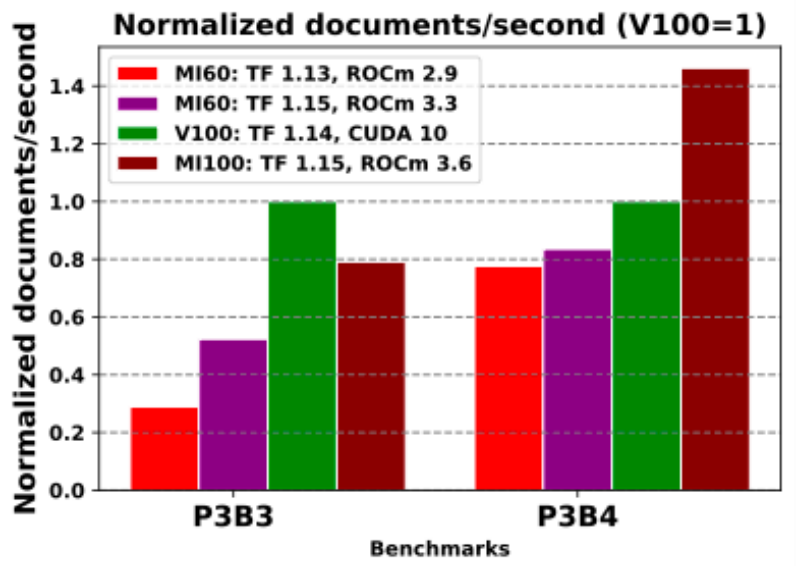


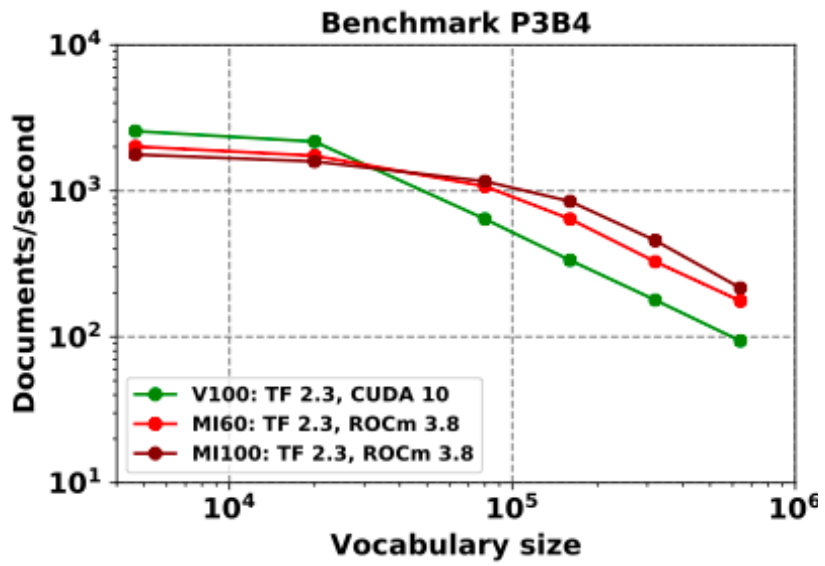
Pilot 3: Population Information Integration, Analysis, and Modeling (2)

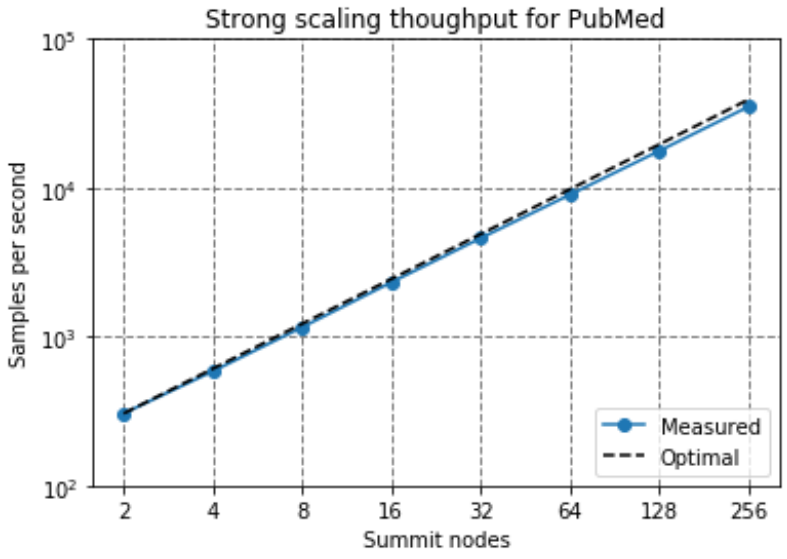
DOE infrastructure critical to the pilot:

- Knowledge Discovery Infrastructure at ORNL - Secure PHI enclave with GPU computing resources for data storage and AI model development
- CITADEL – A privacy-preserving computing solution that enables secure large-scale training and hyperparameter optimization of deep learning models using PHI data on Summit (and future OLCF systems)
- Summit - ALCC allocation for development of large-scale Transformer language models for clinical NLP
- Frontier - Optimizing Transformer pipeline on early access hardware

Impact to exascale mission: Delivering general software modules on Summit and Frontier for large scale (1000+ node) pretraining of Transformer language models for clinical, biomedical, and other domain science applications







IMPROVE: Innovative Methodologies and New Data for Predictive Oncology Model Evaluation (1)

PI: Rick Stevens (ANL)

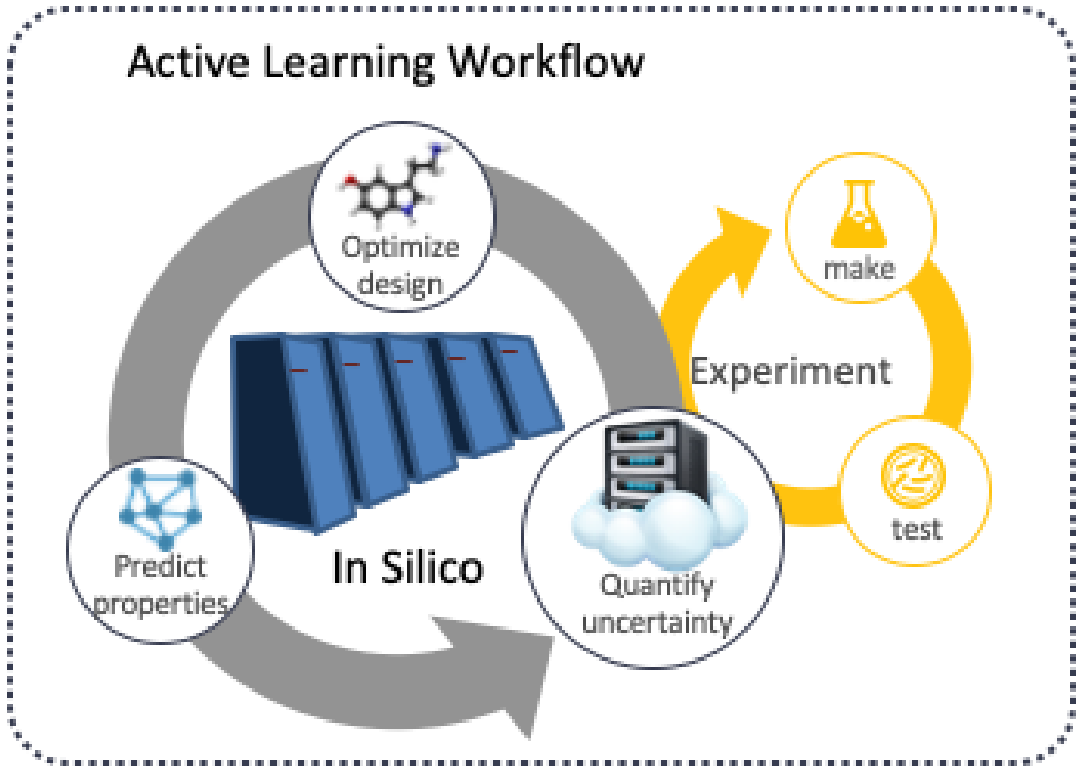
- **Project Goals:**
  1. Development of a robust framework for comparing **deep learning** cancer drug response models
  2. Design and execution of high-throughput experiments aimed at producing new datasets
- **Aim 1: Develop a deep learning model comparison framework**
  - Models will be developed in context generalizability (tumor and drug) and advancing interpretability, explainability, and fairness
  - NCI will fund 3-5 *extramural model design groups* to actively collaborate with ANL in developing and comparing cancer drug-response DL models

* IMPROVE is loosely inspired by the long-standing DOE effort to compare climate model performance supported by BER

IMPROVE: Innovative Methodologies and New Data for Predictive Oncology Model Evaluation (2)

- **Aim 2: Data Generation to Improve Drug Response Models**
  - Development of ”*experimental campaign proposals*” aimed at elucidating what new experimental data are needed to augment existing training data for improving DL drug response prediction models
  - This approach to AI oriented experimental design is also expected to translate to many areas of DOE research and to couple well with research efforts in autonomous discovery
  - ANL will fund one or more *data generation groups* from the public and/or private sectors
- **IMPROVE Anticipated Results:** new datasets optimized for improving ML models, a model comparison framework, a deeper understanding of current landscape of drug response modeling, and new and improved predictive oncology models

Accelerating Therapeutics for Opportunities in Medicine (ATOM)



Goal: Accelerate and democratize drug discovery and optimization

We'll do this by ….

1. Demonstrating acceleration and effectiveness of AI and computing-driven molecular discovery and optimization
2. Developing, validating, and releasing an open platform for molecular discovery and design
3. Applying the platform to making molecules for work on cancer, infectious disease, and underserved needs in the public good

Founding Organizations
October, 2017

Frederick National Laboratory for Cancer Research	UCSF	Lawrence Livermore National Laboratory	gsk
Experimental biology, data analytics, open data and models	Cancer biology and new assay development	High performance computing, data science, and molecular design	Drug discovery, chemistry, and historical dark data

Developed Capabilities

CANDLE: CANcer Distributed Learning Environment
Highlights of Deep Learning Framework

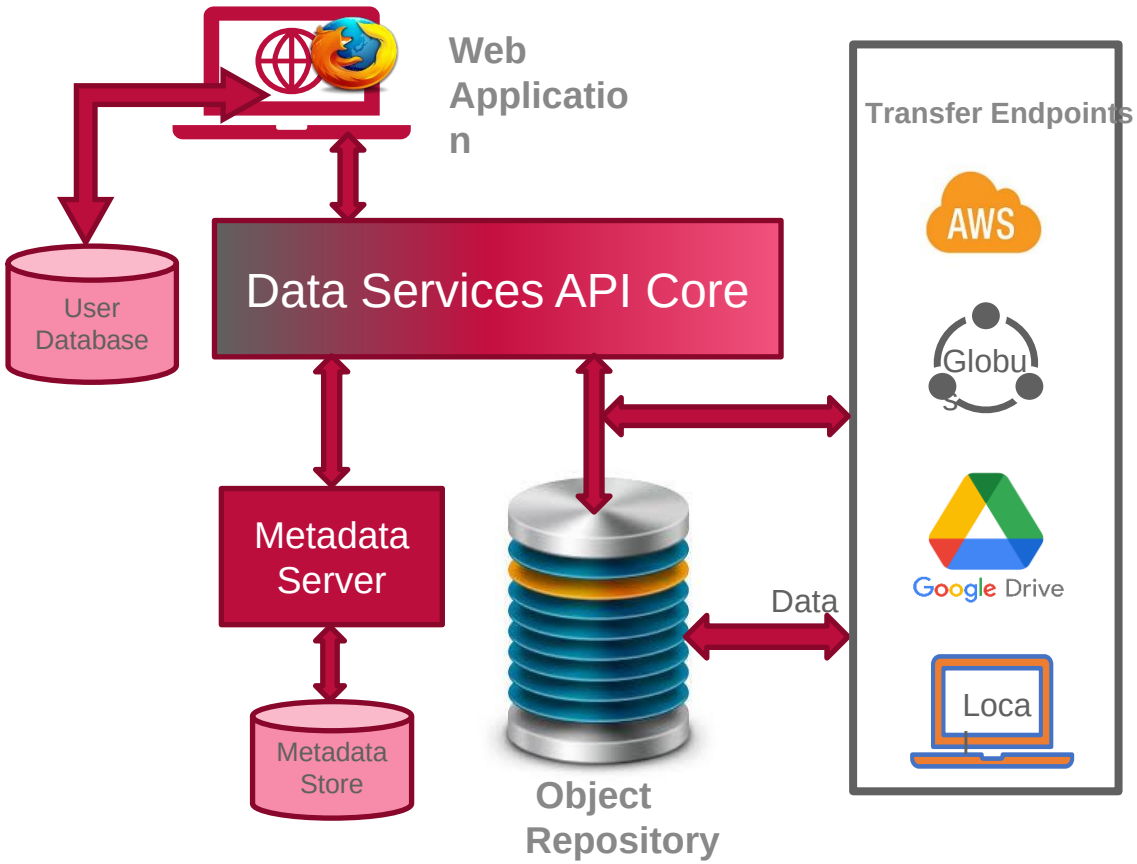
- Open source, Deep Learning software platform accelerates three cancer research challenges:
  - **PILOT 1 - Drug response:** Predictions of tumor response to drug treatments, based on molecular features of tumor cells and drug descriptors
  - **PILOT 2 - RAS pathway:** Identification of key molecular interactions, based on molecular dynamics simulations of proteins, specifically RAS
  - **PILOT 3 - Treatment strategy:** Better characterization of cancer patient trajectories and outcomes using a growing compendium of clinical information
- Scales efficiently on the world's most powerful supercomputers, enabling exascale capabilities
- Current functionalities (on NIH's Biowulf supercomputer):
  - Hyperparameter optimization (HPO) of machine/deep learning models using either grid or Bayesian search
- CANDLE benchmarks are publicly available on Github:
  - **Benchmarks:** https://github.com/ECP-CANDLE/Benchmarks
  - **Documentation:** https://ecp-candle.github.io/Candle/html
  - **FTP site:** https://ftp.mcs.anl.gov/pub/candle/public

At NIH, CANDLE has been used for hyperparameter optimization of models related to analysis of cellular images, e.g., segmentation of subcellular structures such as mitochondria

CANDLE is also being used for Uncertainty Quantification (UQ) for some of the computational models developed in the pilots

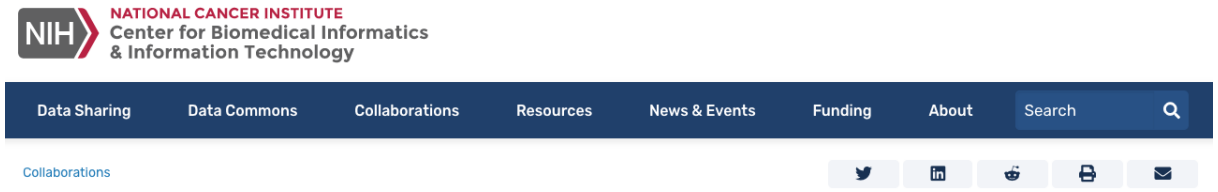
Predictive Oncology Model and Data Clearinghouse (MoDaC)

https://modac.cancer.gov



- **Clearinghouse for annotated mathematical models and data sets from NCI-DOE collaborations, including JDACS4C and ATOM**
- *User-friendly, public facing web interface*
- *Metadata based search capability for locating models and datasets.*
- *Multiple data download options including transfer to AWS S3 bucket.*
- *Public REST API interface*
  - *Programmatic discovery and retrieval of assets.*
  - *Integration with modeling and biomedical analysis platforms.*
- *DOI Support*
  - *Global identifier per asset.*
  - *Shareable link for citations.*

NEW Capabilities for Public Use



https://datascience.cancer.gov/collaborations/nci-doe-capabilities

For more information and technical support, contact: NCI-DOECapabilities@nih.gov

Capability	Type	Impact
Combo: Combination drug response predictor	Model	Enables predictions of drug responses under different experimental configurations.
Uno: Unified drug response predictor	Model	Enables predictions of drug responses under different experimental configurations.
ML-Ready Pathology Reports	Data Set	Provides users with a pathology report data set to use with many of the other capabilities.
MT-CNN	Model	Allows automatic information extraction from free-form pathology report texts. Faster than HiSAN.
HiSAN	Model	Allows automatic information extraction from free-form pathology report texts. More accurate than MT-CNN.
ATOM Modeling PipeLine	Software	Offers an open source, modular, extensible software pipeline for building and sharing models to advance in silico drug discovery.
CANDLE Software Stack	Software	Enables hyperparameter optimization on machine/deep learning models.
NT3: Normal-tumor pair classifier	Model	Classifies RNA-seq gene expression profiles into normal or tumor tissue categories.
TC1: Tissue type classifier	Model	Allows classification of tumor type based on sequence data
DynIm	Software	Enables machine learning-based adaptive multiscale simulations for cancer biology where the input distribution can change over time and the sampling adapts itself to the new distribution.
ANS: Autoencoder Node Saliency	Software	Allows users to understand importance of neural network nodes in autoencoders.
MemSurfer	Software	Computes and analyzes membrane surfaces found in a wide variety of large-scale molecular simulations, enabling assessment of lipid membrane curvature and density.
Active Learning for NLP Systems	Software	Enables rapid annotation of pathology reports via active machine learning framework for natural language processing.

List of Publications on the NCI website

NATIONAL CANCER INSTITUTE
Center for Biomedical Informatics & Information Technology

Data Sharing | Data Commons | Collaborations | Resources | News & Events | Funding | About | Search

Collaborations

NCI-DOE Collaboration Publications

The Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program is the focal point of the strategic, interagency collaboration between the National Cancer Institute (NCI) and the US Department of Energy (DOE) to simultaneously accelerate advances in precision oncology and scientific computing.

Based on a multidisciplinary, team science approach, JDACS4C's three research pilots were co-designed by NCI and DOE; align with several existing NCI and DOE programs; and are jointly led by DOE and NCI supported scientists. These teams include scientists from NCI and Frederick National Laboratory for Cancer Research and experts from DOE national laboratories: principally Argonne, Lawrence Livermore, Los Alamos, and Oak Ridge.

Below is a list of the publications. For more information on the NCI-DOE Collaboration, visit the JDACS4C page.

https://datascience.cancer.gov/collaborations/nci-doe-publications

For more information contact: NCI-DOECapabilities@nih.gov

Select publications

2020, DeepFreeze: Towards Scalable Asynchronous Checkpointing of Deep Learning Models, 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing

2020, Distributed Bayesian Optimization of Deep Reinforcement Learning Algorithms, Journal of Parallel and Distributed Computing

2020, Generation and Evaluation of Synthetic Patient Data, BMC Medical Research Methodology

2020, How Anionic Lipids Affect Spatiotemporal Properties of KRAS4B on Model Membranes, The Journal of Physical Chemistry B

2020, Knowledge Graph-Enabled Cancer Data Analytics, IEEE Transactions on Emerging Topics in Computing

2020, Presence or Absence of Ras Dimerization Shows Distinct Kinetic Signature in Ras-Raf Interaction, Biophysical Journal

2020, Privacy-Preserving Deep Learning NLP Models for Cancer Registries, IEEE Transactions on Emerging Topics in Computing

2020, Selective Information Extraction Strategies for Cancer Pathology Reports with Convolutional Neural Networks, Recent Advances in Big Data and Deep Learning

2020, The Plasma Membrane as a Competitive Inhibitor and Positive Allosteric Modulator of KRas4B Signaling, Biophysical Journal

2020, Using Case-level Context to Classify Cancer Pathology Reports, PLOS One

2019, A Knowledge Graph Approach for the Secondary Use of Cancer Registry Data, 2019 IEEE EMBS International Conference on Biomedical & Health Informatics

2019, A Massively Parallel Infrastructure for Adaptive Multiscale Simulations: Modeling RAS Initiation Pathway for Cancer, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis

2019, Adversarial Training for Privacy-Preserving Deep Learning Model Distribution, 2019 IEEE International Conference on Big Data (Big Data)

2019, AI Meets Exascale Computing: Advancing Cancer Research With Large-Scale High Performance Computing, Frontiers in Oncology

2019, Autoencoder Node Saliency: Selecting Relevant Latent Representations, Pattern Recognition

2019, Classifying Cancer Pathology Reports with Hierarchical Self-attention Networks, Artificial Intelligence in Medicine

2019, Combating Label Noise in Deep Learning Using Abstention, Proceedings of the 36th International Conference on Machine Learning

2019, Computationally Efficient Learning of Quality Controlled Word Embeddings for Natural Language Processing, 2019 IEEE Computer Society Annual Symposium on VLSI

2019, Computing Long Time Scale Biomolecular Dynamics Using Quasi-stationary Distribution Kinetic Monte Carlo (QSD-KMC), The Journal of Chemical Physics

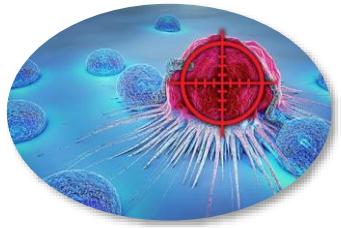
New Opportunities

FY22 and beyond

Envisioning Computational Innovations for Cancer Challenges (ECICC)
Community Driven Collaborative and Team Science Activities

ECICC Scoping Meeting
March 2019

>70 computational and cancer scientists;
Four challenge areas:
1) Digital twin,
2) Adaptive treatment,
3) Synthetic data,
4) ML-driven hypothesis generation

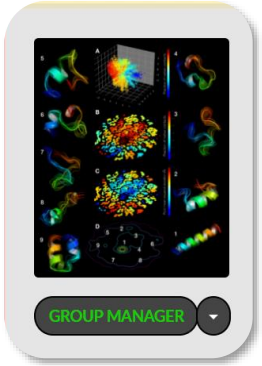


Cancer Patient Digital Twin Ideas Lab
July 2020

5-day virtual event
130 applicants
30 attendees
6 experienced mentors
6-month initial seed projects

ECICC Community Site
Created 2019

Over 250 members
Ongoing community platform
Visit website at
https://ncihub.org/groups/cicc



Predictive Rad Onc Workshops
February 2021

Joint NCI and DOE organizing committee
Extramural PI steering committee
Explore frontiers of precision radiation therapy

https://events.cancer.gov/cbiit/radonc2021

Thank you

Questions & comments?