Impacting Precision Medicine with DOE High Performance Computing, Machine Learning and Artificial Intelligence Research

Kerstin Kleese van Dam
Ji-Hwan Park, Shantenu Jha, Frank Alexander, Shinjae Yoo
Brookhaven National Laboratory
Computational Science Initiative
kleese@bnl.gov

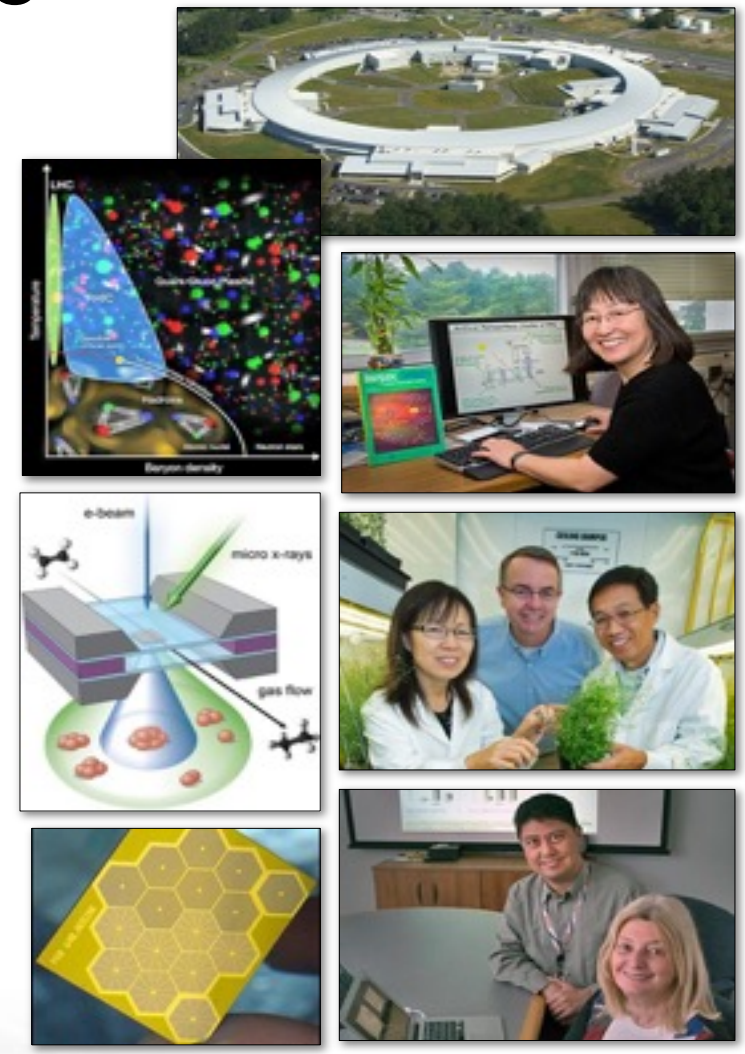
BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

Brookhaven at a Glance

- Established in 1947
- One of 17 DOE Labs
- Based on Long Island, NY
- Seven Nobel Prizes
- 2400 employees, ~400 Grad/Undergrad students
- Six DOE user facilities
- 3198 Facility Users, 2176 Visiting Scientists
- 5322 acres, 316 Buildings
- **Discovery science and transformative technology that power and secure the nation's future**
- **Nuclear and High Energy Physics, Materials and Chemical Sciences, Data Science**



Brookhaven Operates Many Data-rich Experimental Facilities

- Relativistic Heavy Ion Collider (**RHIC**)
- National Synchrotron Light Source II (**NSLS-II**)
- Center for Functional Nanomaterials (**CFN**)
- Accelerator Test Facility (**ATF**)
- Large Hadron Collider (LHC) **ATLAS**
- Atmospheric Radiation Measurement (**ARM**) program
- **Belle II**: computing for neutrino experiment
- Quantum chromodynamics (**QCD**) computing facilities for Brookhaven, RIKEN and U.S. QCD communities
- Brookhaven Linac Isotope Producer (**BLIP**)
- NASA Space Radiation Laboratory (**NSRL**)
- **Coming - CryoEM Facility**

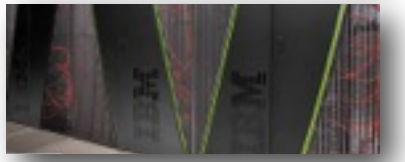
RHIC

NSLS II

CFN

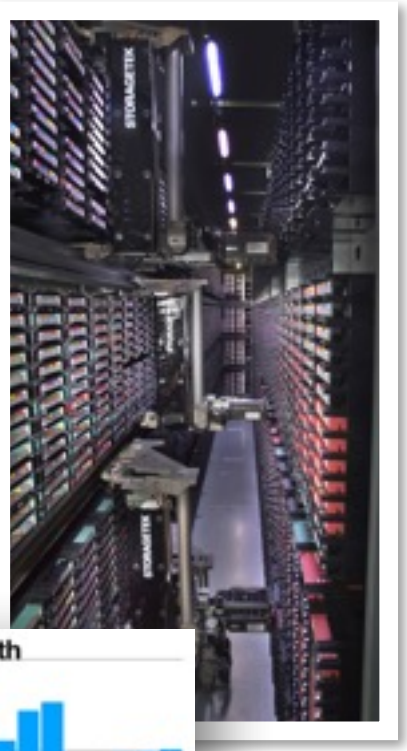
ATLAS

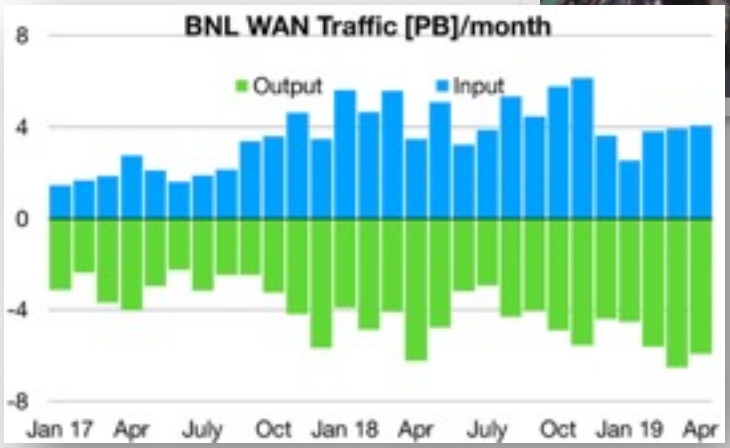
QCD



Brookhaven Lab Data by the Numbers

- **One of the top-five largest scientific archives in the world***
  - **~160 PB** of data **archived** to date
  - **32 million files archived** (**26 PB**)
  - **16 million files restored** (**29 PB**)
- **800 PB** of data **analyzed** in FY18
- **~110 PB** of data **transferred** annually
  - Data import: 52 PB
  - Data export: 57 PB
  - ~30% increase from previous year



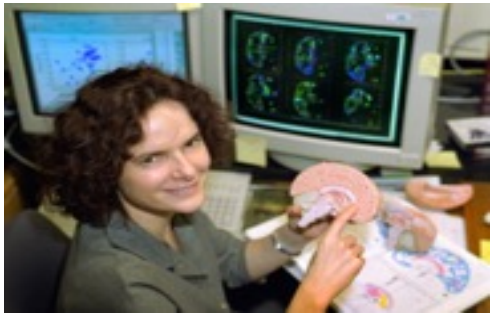


*Source: http://www.hpss-collaboration.org/customersT.shtml (as of 06/2019)

History in Medical Innovation

- Patented easy-to-use kit that attaches technetium-99m to red blood cells, so doctors can see blood movement.
- L-dopa for **Parkinson's disease treatment** evolved from Brookhaven study of relationship between trace elements and neurological diseases.
- Thallium-201, a radioisotope developed at Brookhaven, is used in **heart stress** tests worldwide.
- Pioneered positron emission tomography "**PET**" **scan** technology.
- Tin-117m DPTA is used in heavy sedation for **bone cancer** patients, providing extreme pain relief to sufferers.

SynchroPET Inc.



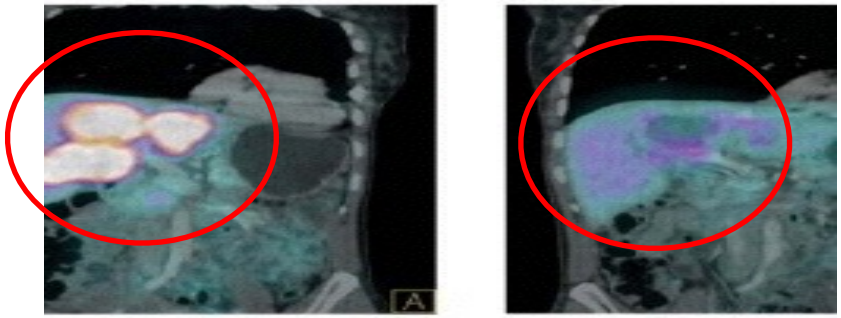
Nora Volkow was Brookhaven's former associate director for Life Sciences who spent over 15 years at the Lab using PET technology to investigate the physical causes of addiction in the human brain.

Using Today's Research for Medical Progress

Brookhaven Linac Isotope Producer (BLIP)

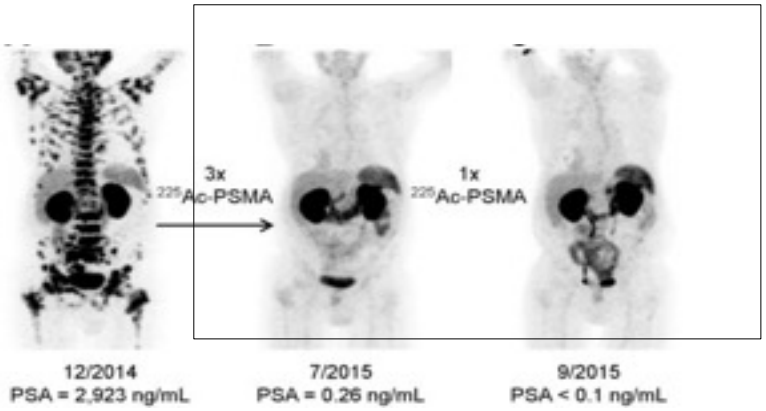
- Production and development of medical isotopes for diagnostics and therapy
- Bi-213 successfully demonstrated for lung cancer treatment
- R&D to produce Ac-225: highly promising alpha emitter for cancer treatment

BLIP



Liver tumor before treatment

After Bi-213 treatment



Prostate cancertherapy
J. Nucl. Med., **57**(12) 1941 (2016)

Pivoting Brookhaven Lab Machine Learning, Artificial Intelligence and High Performance Computing Research Toward Precision Medicine

Key Challenge Examples at Brookhaven's Large-scale Facilities

Real-time Analysis and Steering of Experiments–Challenges:

- CFN – 400 images/sec
- NSLS II – up to 5 TB/s in burst

Extreme Scale Data Management and Analysis:

- **800 PB of data analyzed in FY18**
- Moved data to compute and storage in 2016 – 1.6 exabyte

CSI Research Focus Areas

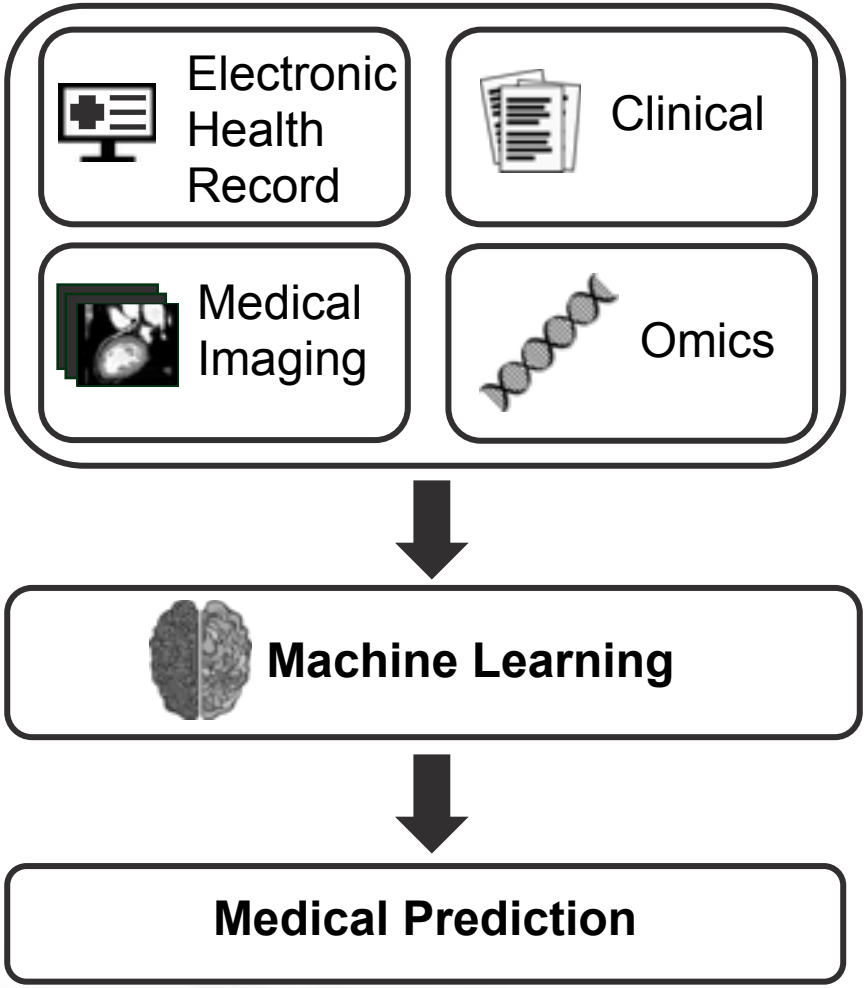
- **Applied Mathematics, Artificial Intelligence and Machine Learning for Real-time Data Analysis**
- Programming Models, Compilers, and **Resource Management for AI- and ML-based Workflows**
- Specialized Systems and Architectures for AI and ML Real-time Analysis Applications
- Large scale Numerical Modeling for Chemistry, Materials, QCD and Design of Experimental Facilities
- Extreme Scale Operational Data and Computing Facilities

Machine Learning for Medicine

NHIS Ilsan Hospital
- Dementia
- Depression

Veterans Affairs
- Prostate Cancer
- Cardiovascular Disease
- Suicide Prevention

NIH National Cancer Institute
- Computational methods
- Large-scale computation
- Impact of existing therapies



Accurate Prediction of Alzheimer's Disease with Novel Machine Learning

- Brookhaven Lab, Columbia University and the National Health Insurance Service (NHIS) Ilsan Hospital and South Korea collaborated to develop Alzheimer Prediction methods.

- Brookhaven Lab has developed advanced, highly scalable machine learning algorithms for accurately predicting Alzheimer's disease.

- Algorithms were trained on **magnetic resonance imaging and brain phenotyping clinical data** to make predictions of diagnosis

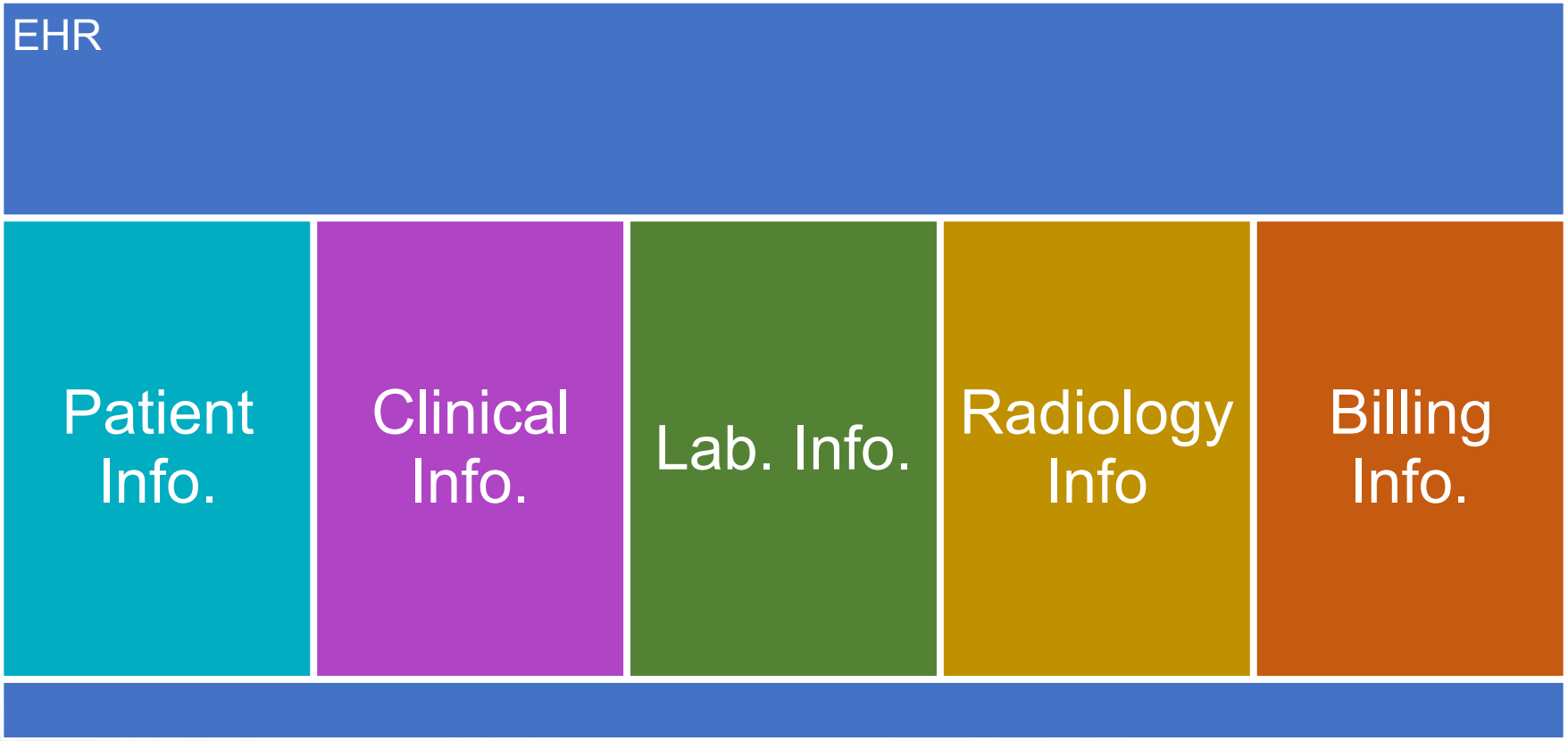
- **Achieved nearly 100% Alzheimer's detection accuracy and 83% prediction accuracy for early-stage Alzheimer's**.

Affordable EHR for Screening Alzheimer's disease (AD)

- Biomakers - the collection of bio-specimen (e.g., serum or fluid) or imaging data
  - ➔ Time consuming

- Electronic health records (EHR)
  - **not require additional time or effort** for data collection
  - Increase the size of EHR data due to digitalization

Overview of EHR



A few predefined features

- In prior work, predefined features
  - sociodemographic (age, sex, education)
  - lifestyle (physical activity)
  - midlife health risk factors (systolic blood pressure, BMI and total cholesterol level)
  - cognitive profiles

- **Multi-factor** models best predict risk for dementia

➔ **Machine learning**

Machine learning on high-dimensional EHR

- Use a large nationally representative (South Korea) sample cohort
- Construct and validate data-driven **machine learning** models to predict future incidence of AD using the extensive measures collected within high-dimensional EHR
- Demonstrate the feasibility of developing accurate prediction models for AD

Korean EHR data

- Korean National Health Insurance Service - National Elderly cohort Database
- 6,435 features
- 430,133 individuals (> 65 yrs, 10% sample of randomly selected elderly individuals)
- 2002 – 2010, South Korea

High-dimensional Features

National Elderly cohort Database (DB)

Health Screening (HS) DB	Participant Insurance Eligibility (PIE) DB	Healthcare Utilization (HU) DB
21 Features: laboratory values, health profiles, history of family illness	2 Features: sex, age	6,412 features including ICD-10 codes and medication codes

Machine learning analysis

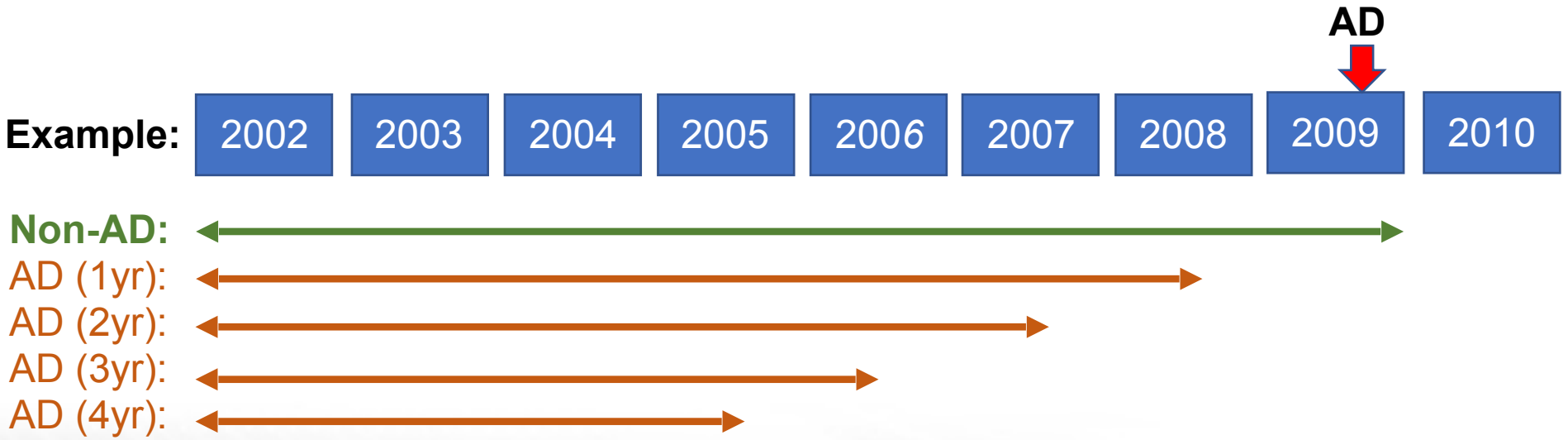
- Input: High-dimensional EHR data
- Methods
  - Random forest, support vector machine (SVM), logistic regression
- Task: Can machine learning be used to predict future incidence of Alzheimer's disease using electronic health records?

Definition of data

- Two criteria
  - (Korean) ICD-10 code:
    - Dementia in AD - F00, F00.0, F00.1, F00.2, F00.9
    - AD - G30, G30.0, G30.1, G30.8, G30.9
  - Dementia medication: e.g., donepezil, rivastigmine, galantamine, and memantine
- ***Definite AD**: ICD-10 code + medication*
- ***Probable AD**: only ICD-10*

Data range for n-year prediction

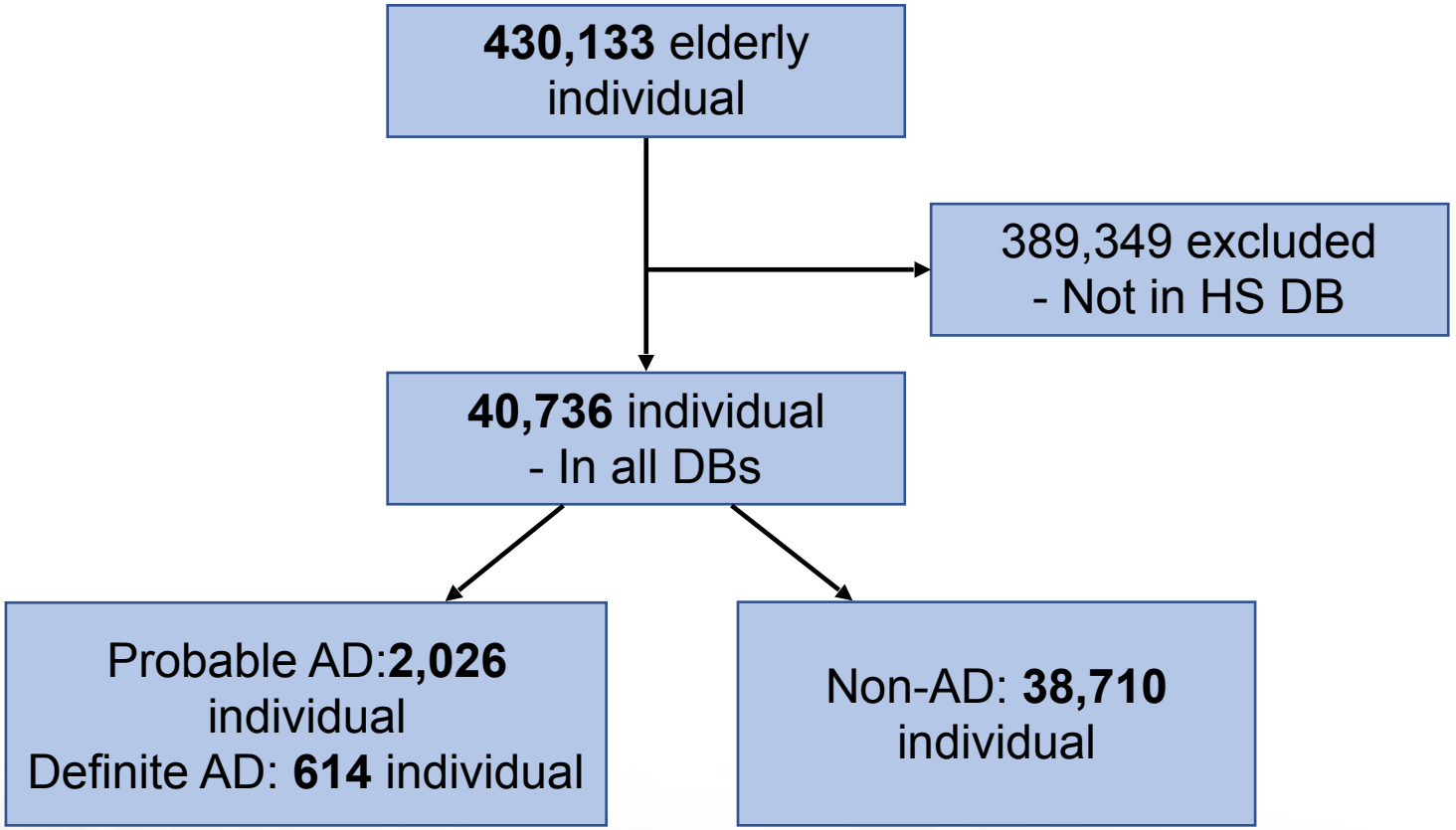
- AD group: between 2002 and the year of incident AD – *n*
- Non-AD group: 2002 to 2010 – *n*



Data Preprocessing

- EHR alignment
- ICD-10 and medication coding
  - the first disease category codes: e.g., **F00**.0
  - the first 4 characters for the medication codes representing main ingredients: e.g., **1498**01ATB
- Rare disease exclusion (≤ 5)
- Records exist in all the three databases (HS, PIE,HU)

of data samples

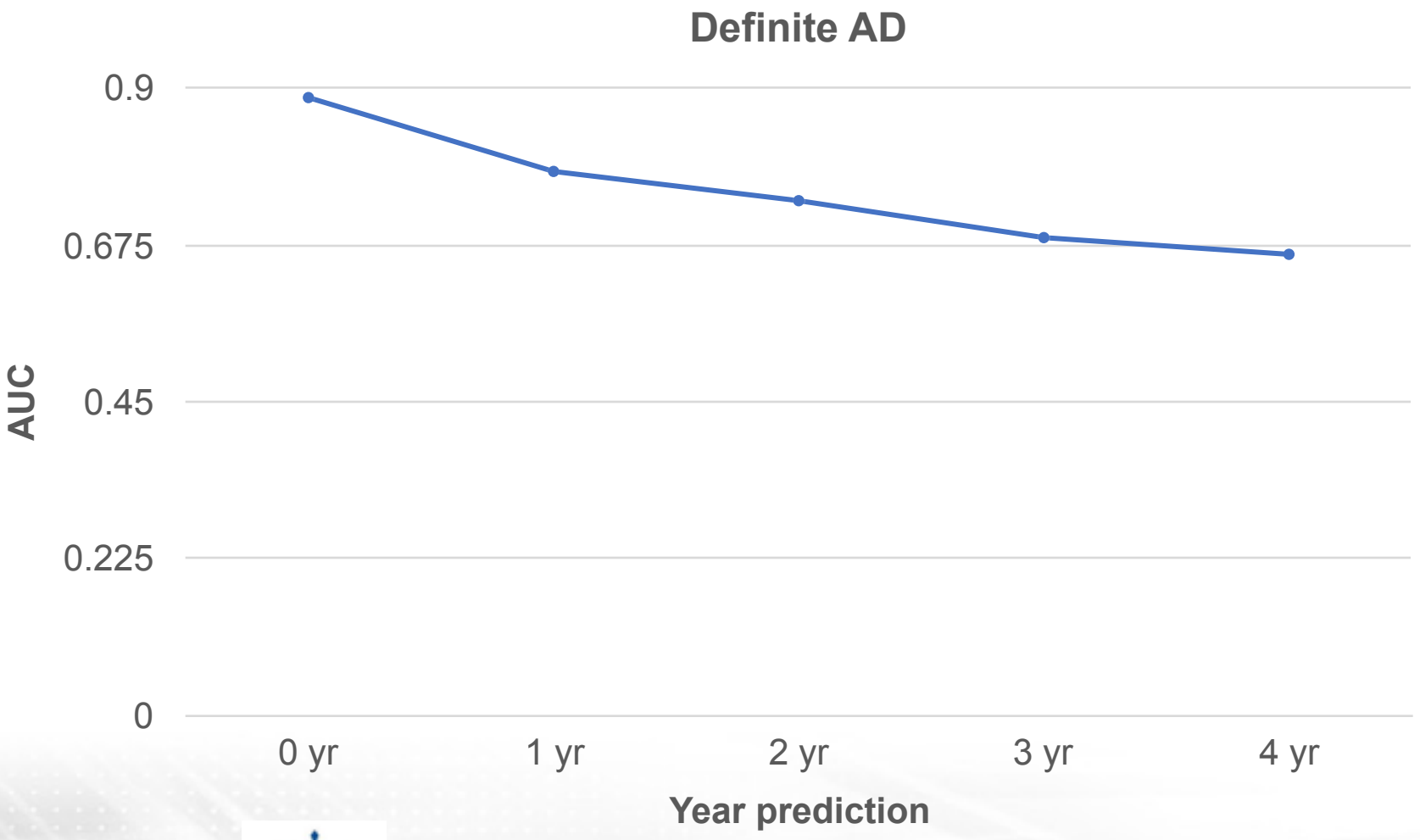


Sample characteristics

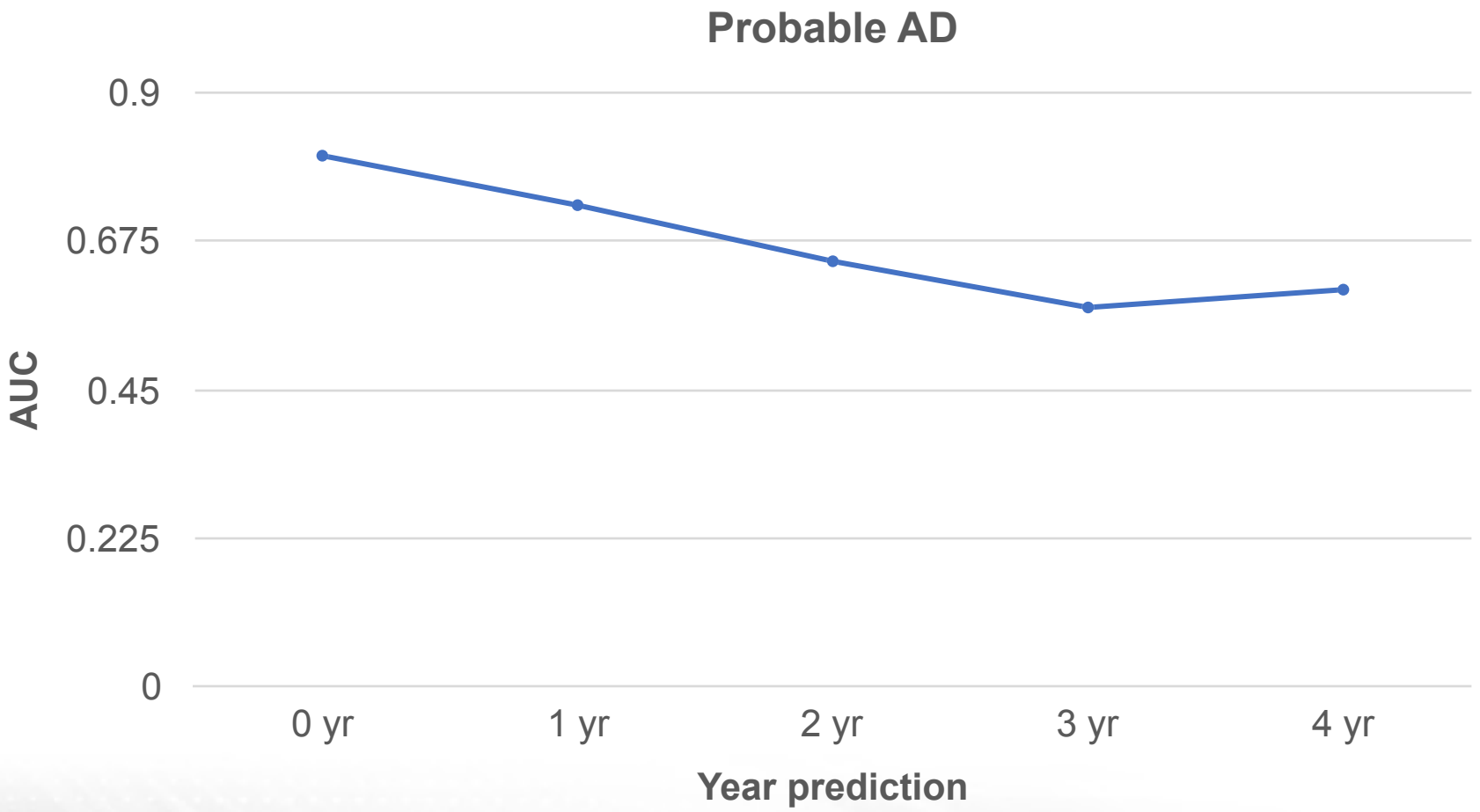
	Definite AD	Probable AD	Non-AD
Number	614	2,026	38,710
Income	$ 60k ($57.3k-$62.7k)	$59k ($58.7k-$59.3k)	$60.2k ($58.7k-$61.7k)
Age	80.67 (80.2-81.1)	79.2 (79.0-79.5)	74.5 (74.4-74.5)
sex	Male:229 (37%) Female:285 (63%)	Male:733 (36%) Female:1,293 (64%)	Male:18,200 (47%) Female:20,510 (53%)

*Based on the 0-year prediction model.

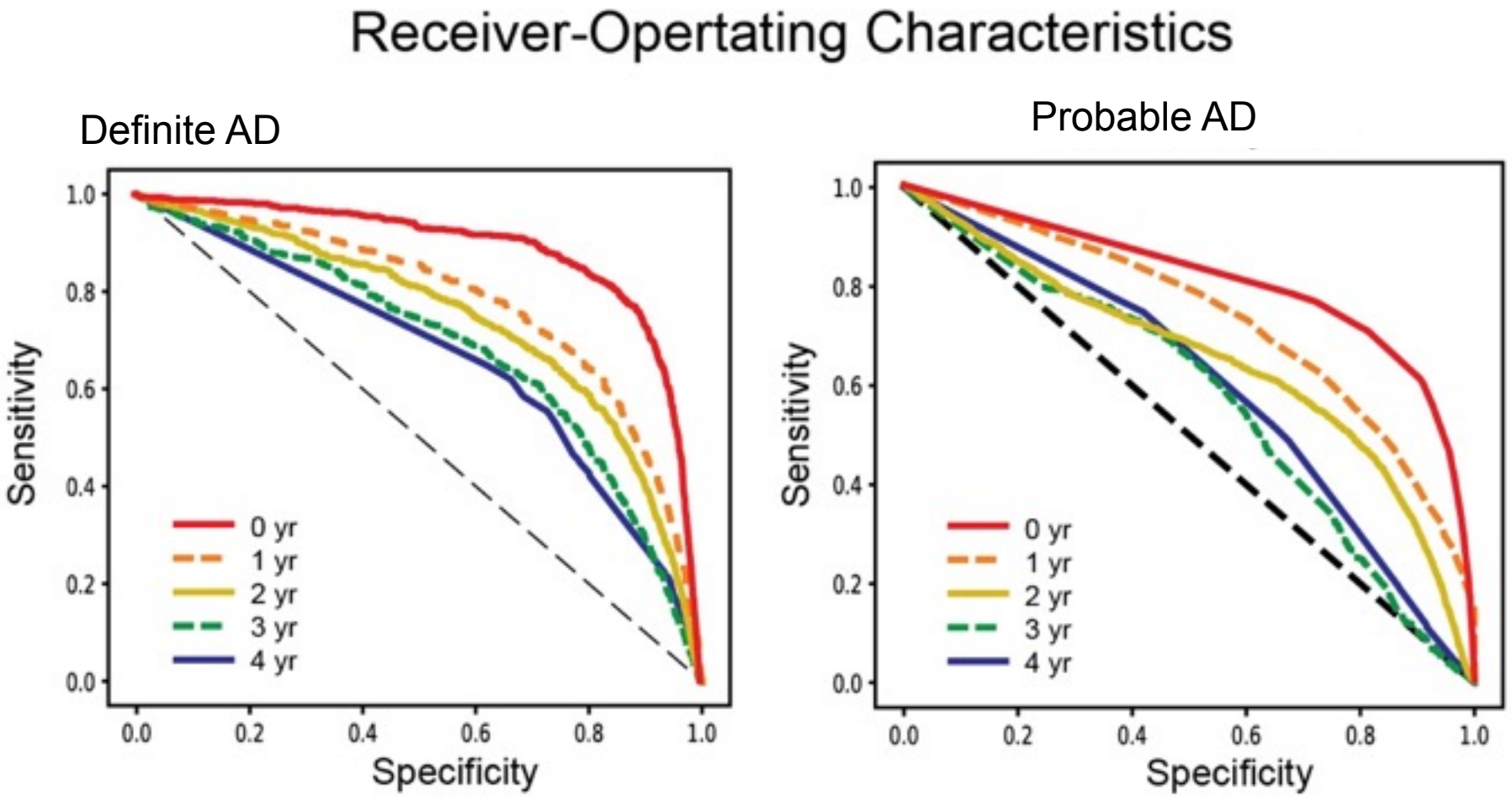
N-year prediction for definite AD



N-year prediction for probable AD



Model prediction result - ROC



Important features

Name	b value
Hemoglobin (H)	-0.902
Age (Demo)	0.689
Urine protein (H)	0.303
Zotepine (antipsychotic drug) (M)	0.303
Nicametate Citrate (vasodilator) (M)	-0.297
Other degenerative disorders of nervous system in diseases classified elsewhere (D)	-0.292
Disorders of external ear in diseases classified elsewhere (D)	0.274
Tolfenamic acid 200mg (pain killer) (M)	0.266
Adult respiratory distress syndrome (D)	-0.259
Eperisone Hydrochloride (antispasmodic drug) (M)	0.255

(H): Health checkup
(M): Medication
(Demo): Demographics
(D): Disease

Summary (1)

- Our model AUC: **0.887** (0yr), **0.781** (1yr), **0.662** (4yr)
- Prior models AUC: 0.5 ~ 0.78

- Detected interesting EHR-based features associated with incident AD

Summary (2)

- Presents the first data in predicting future incident AD using **data-driven machine learning** based on **large-scale EHR**

- Support to the development of **EHR-based AD risk prediction** that may enable **better selection of individuals at risk for AD** in clinical trials or early detection in clinical settings

Future work

- Generalize our findings to ethnicities other than Korean or to different healthcare systems

- Apply deep neural networks such as a recurrent neural network (RNN)

Machine-Learning-enabled Prostate Cancer Occurrence and Progression Prediction

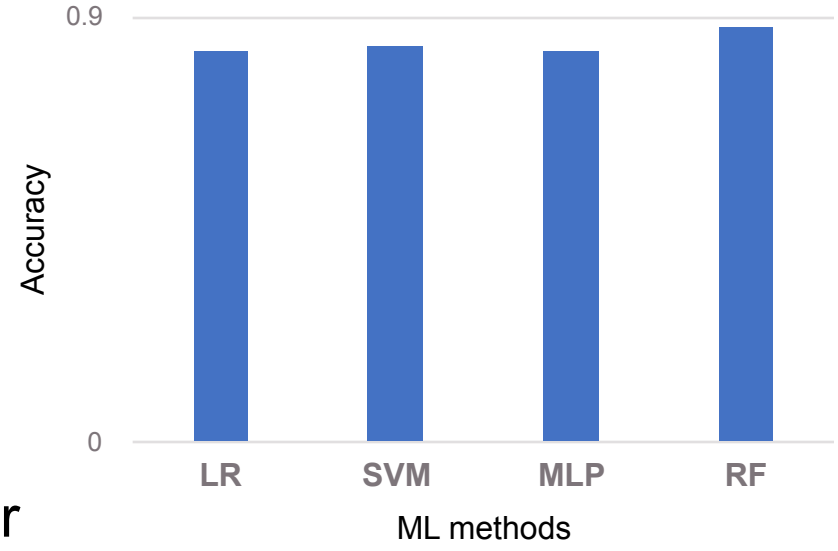
- Preliminary study to accurately predict prostate cancer occurrence in 1-2 years and forecasting of prognosis for prostate cancer patients.

- **We can successfully perform tasks related to prostate cancer using only Electronic Health Records (EHR) data, and the approach also can be applied to other types of cancer**.

- Exploited patients' demographics, disease and medication histories and physical examination records to predict prostate cancer.

Machine-Learning-enabled Prostate Cancer Occurrence and Progression Prediction (2)

- Comparison of machine learning algorithms for predicting prostate cancer occurrence in 1 or 2 years.
- Four algorithms were compared:
  - Logistic regression (LR)
  - Support vector machine (SVM)
  - Multilayer perceptron (MLP)
  - Random forest (RF).
- Achieved high prediction accuracy in all algorithms without using other types of data. The result showed that our task is easily verifiable and robust.



AEOLUS: Advances in Experimental Design, Optimization, and Learning for Uncertain Complex Systems

AEOLUS Motivation: Translation Tools to Precision Medicine

- Remarkable advances in simulation of large-scale complex DOE-relevant systems on leading-edge supercomputers (forward problem)
- Central goal: To fully realize the power of scientific simulation as a basis for scientific discovery, technological innovation, and rational decision-making, it is **imperative to look beyond the forward problem to tackle the outer loop of optimization** for:
  - learning mechanistic predictive models from data
  - optimal experimental design
  - optimal control and design

…all under uncertainty

- Transforming DOE's simulation capabilities into capabilities for optimizing learning and control/design under uncertainty will maximize the impact of DOE investments by augmenting the information learned from expensive computer and laboratory experiments and enhancing predictive medicine performance

Grand Challenge: Transform DOE's Simulation Capabilities to Optimize Learning and Control Under Uncertainty in Precision Medicine

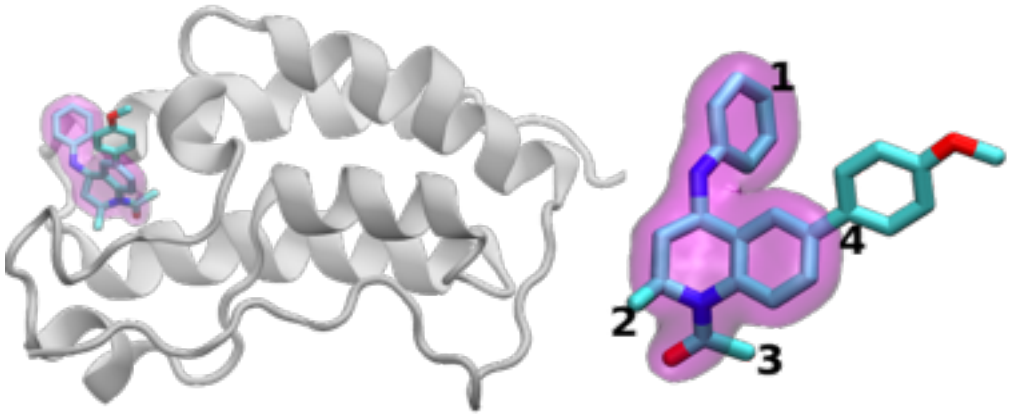
- Despite a long history, **optimization for learning and control/design under uncertainty remains intractable with conventional methods.**
- Forward problems that govern the inner loops of the optimization problems are expensive to execute (due to severe nonlinearity, heterogeneity, multiphysics/ multiscale coupling), often requiring long run times on leading-edge HPC systems.
- Optimization variable spaces, representing design, control, inference, and experimental design variables, are high dimensional, stemming from discretizations of infinite dimensional fields (e.g., initial/ boundary conditions, sources, geometry, coefficients, observation operators).
- Target simulation problems are characterized by high-dimensional uncertain parameters, often from discretization of infinite dimensional fields.

Drug Resistance: Rapid and Accurate Binding Affinity

- Binding affinity/binding free energy (BFE):
  - Predicts measure of goodness of compounds binding with their target proteins.
  - Free energy calculation using ensembles of molecular dynamics simulations
- Given large span of possibilities, determine effective way to screen candidates before more accurate methods to determine in “realistic” discovery situation (time to solution, error, etc.)

Use Case Study

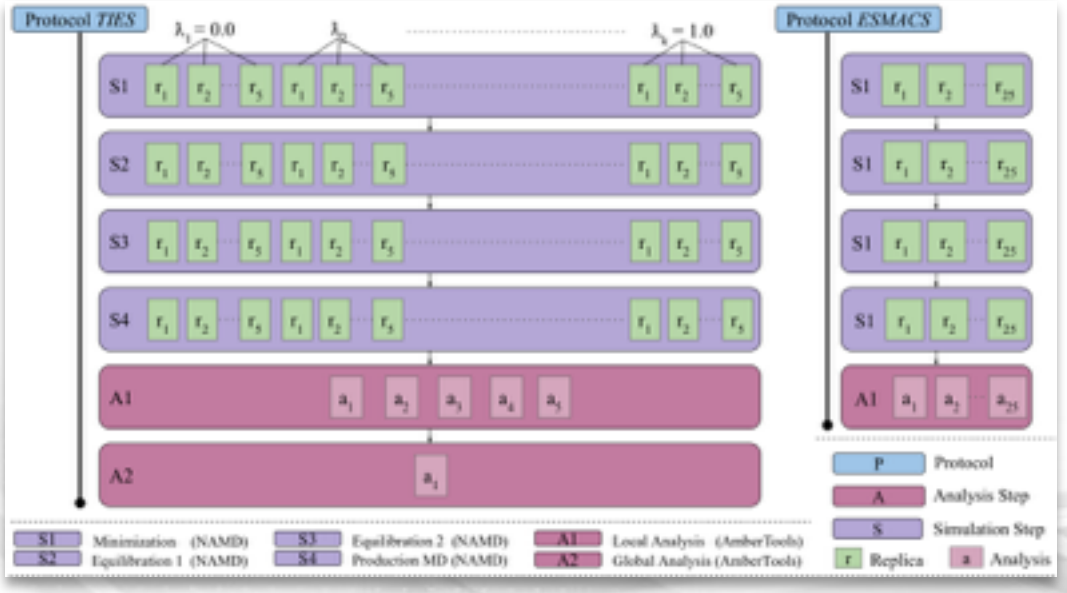
- Original study GSK and UCL
  - 16 drug candidates, BRD4 inhibitor
- Non-adaptive implementation of ESMACS and TIES protocols at scale on average:
  - ESMACS 10K core-hours
  - TIES 25K core-hours
- Millions of candidates to “hits” to leads
- Typical lead optimization involves 10,000 small molecule (drug) candidates
  - 250 million core hours
- Without **specialized tools using building blocks**, otherwise sequential and separate



- Congeneric series of small molecule inhibitors binding to BRD4-BD1
- Blind test of ligands, collaboration with GSK.

A Tale of Two Protocols: Trading Off Accuracy and Cost

- **TIES** (alchemical **protocol**) employs enhanced sampling at each lambda window to compute relative binding affinities.
- **ESMACS** (endpoint **protocol**) is a computationally cheaper but less rigorous method used to directly compute the binding strength of a drug.
- **TIES** 2.5 – 3x cost **ESMACS**.

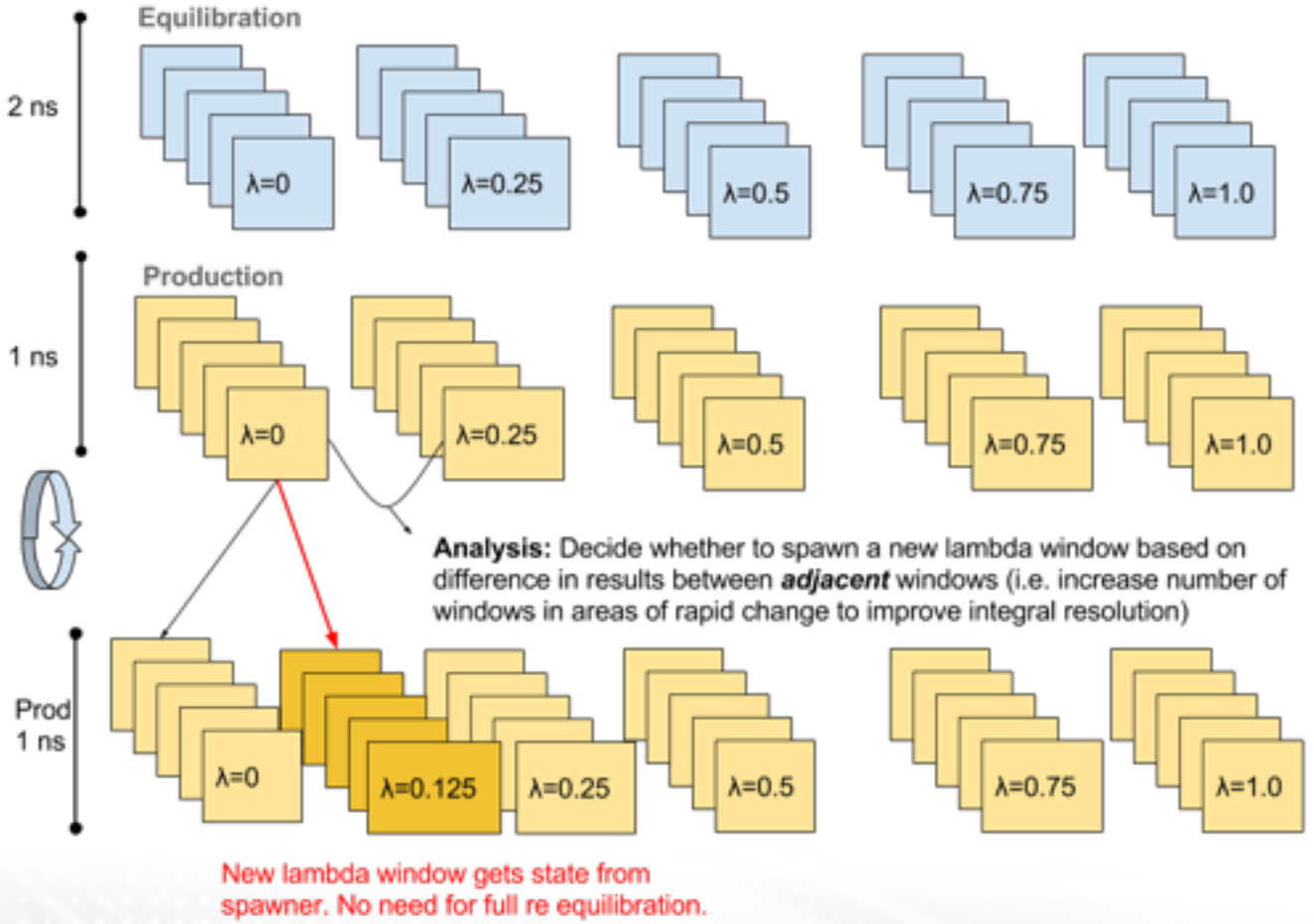


TIES Adaptive Simulation Methods

- Objective: Reduce cost to completion (CPU-hours) while maintaining or reducing errors (kcal/mol)
- Two adaptive methods using **TIES**:
- ***Adaptive Quadrature:***
  - ***Lambda values assigned runtime***
  - ***Traditional approach: Lambda values assigned a priori***

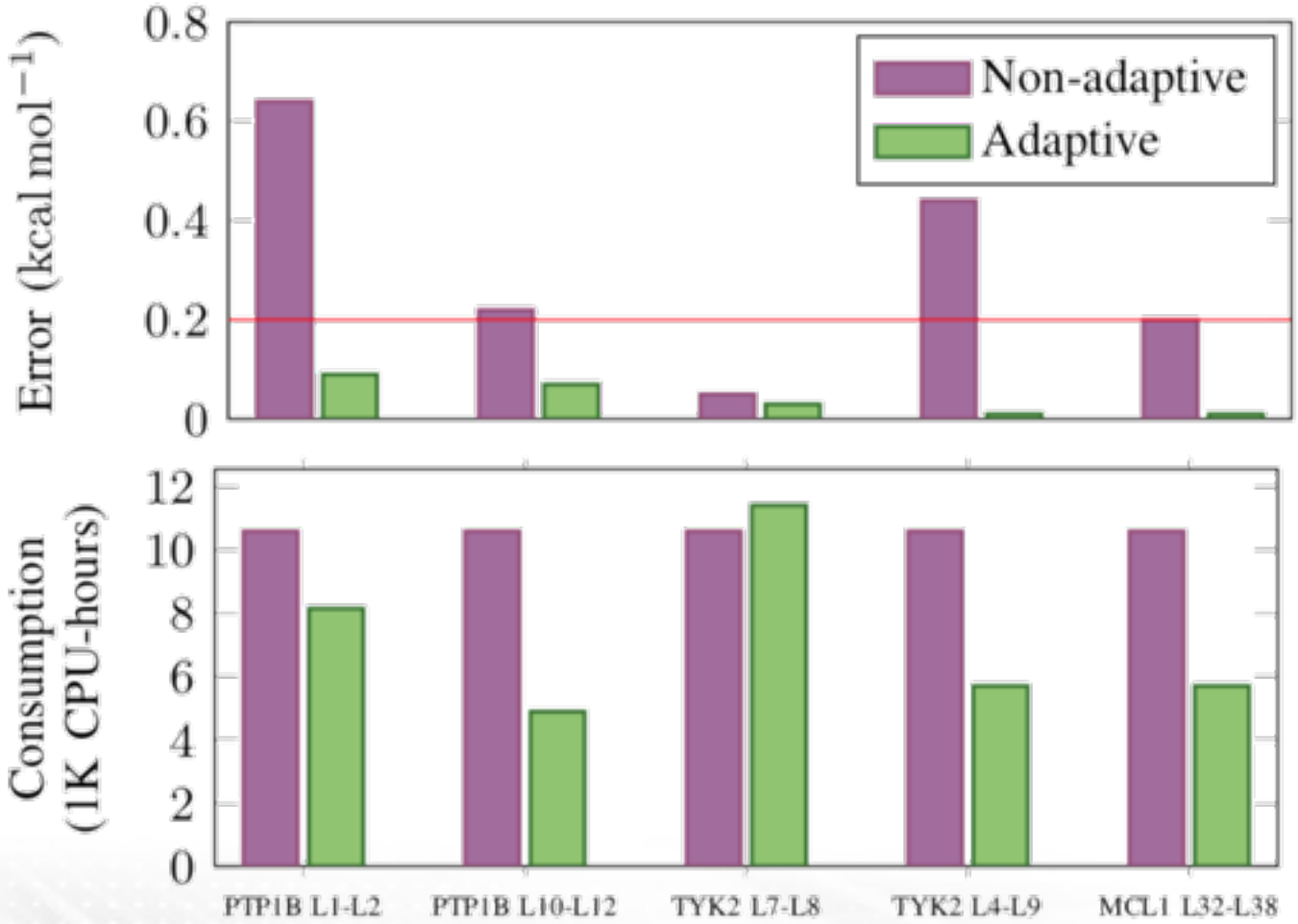
- ***Adaptive Termination***
  - ***Terminates simulations during runtime upon reaching converge criteria (ΔG)***
  - ***Traditional approach: full duration assigned a priori***

TIES Adaptive Simulation Methods

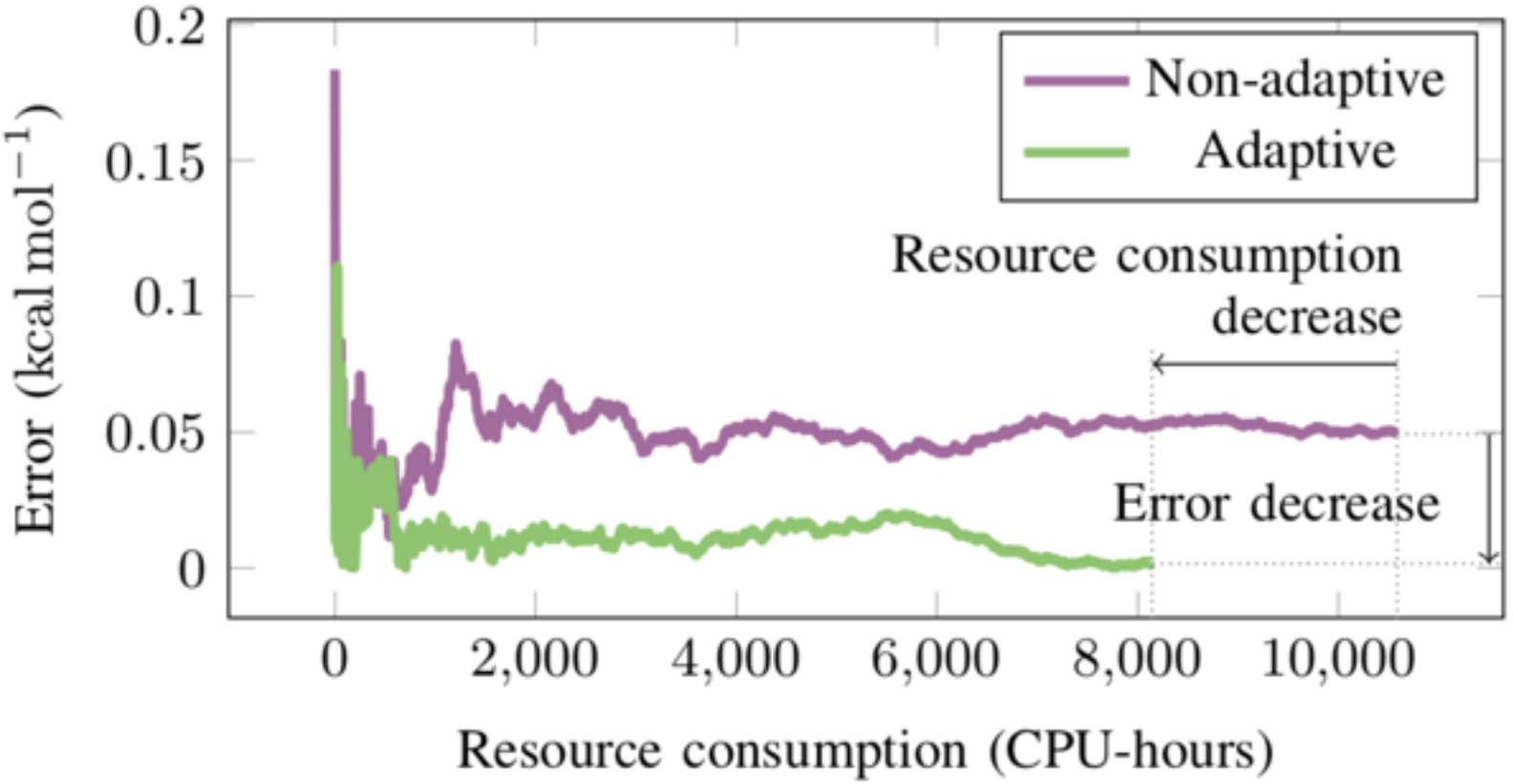


All simulations finish at the same time, and resources (number of cores) are redistributed based on how many windows are being sampled in each step.

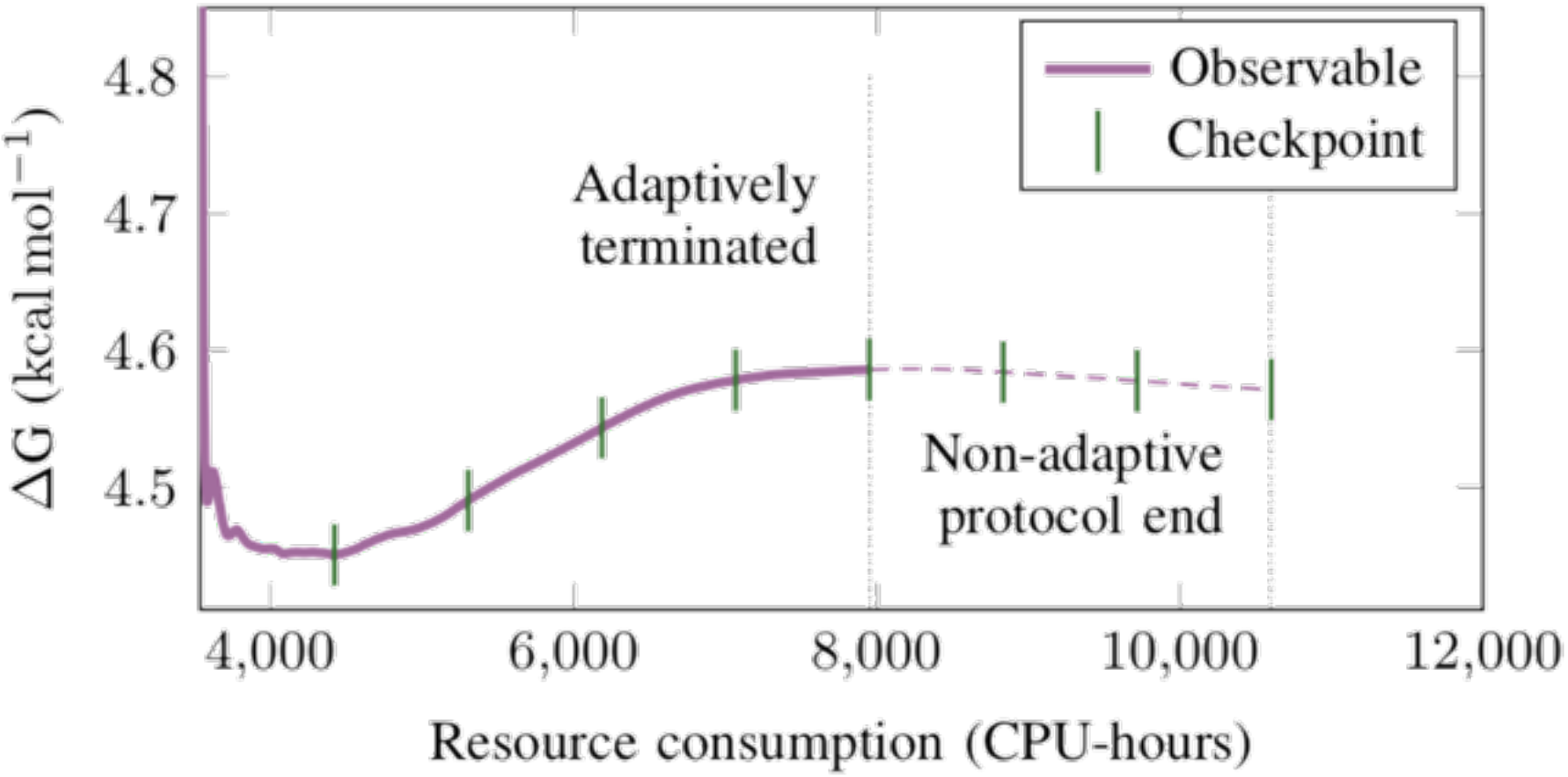
TIES Adaptive Quadrature Results



TIES Adaptive Quadrature Results



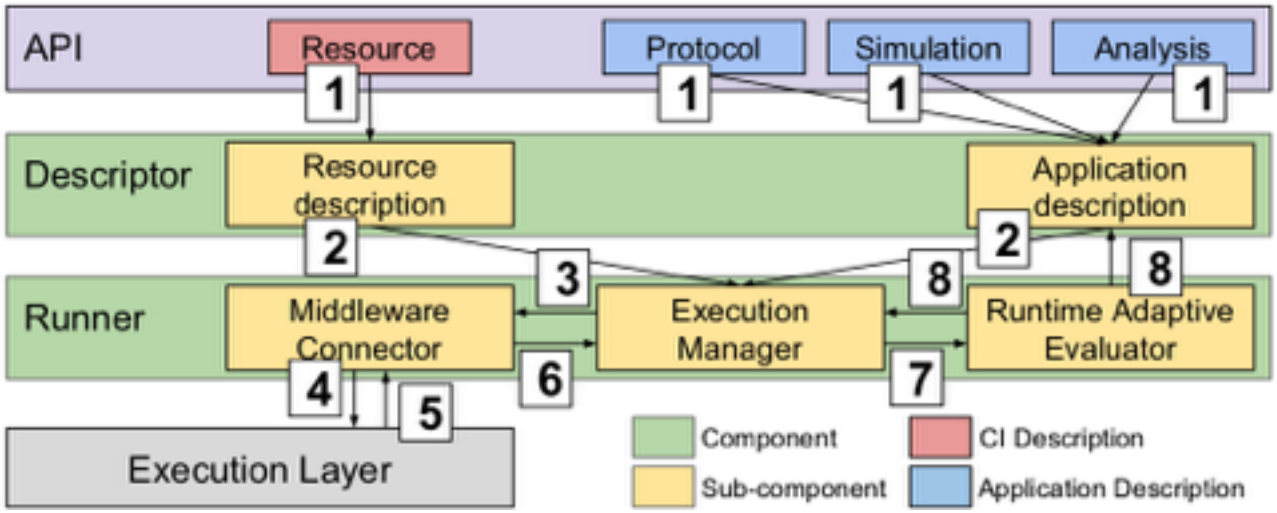
TIES Adaptive Termination Results



HTBAC: High-throughput Binding Affinity Calculator

- Python library for defining and executing ensemble-based biosimulation protocols
  - Protocols expressed and implemented using HTBAC's API.
  - HTBAC uses RADICAL-Cybertools (RCT): Ensemble Toolkit (**EnTK)** and RADICAL-Pilot (**RP)**.
- Implemented ESMACS and TIES protocols
- Define additional adaptivity parameters that are passed down to the underlying runtime system.
- **Goal - Further savings through optimized resource management and usage on HPC system.**

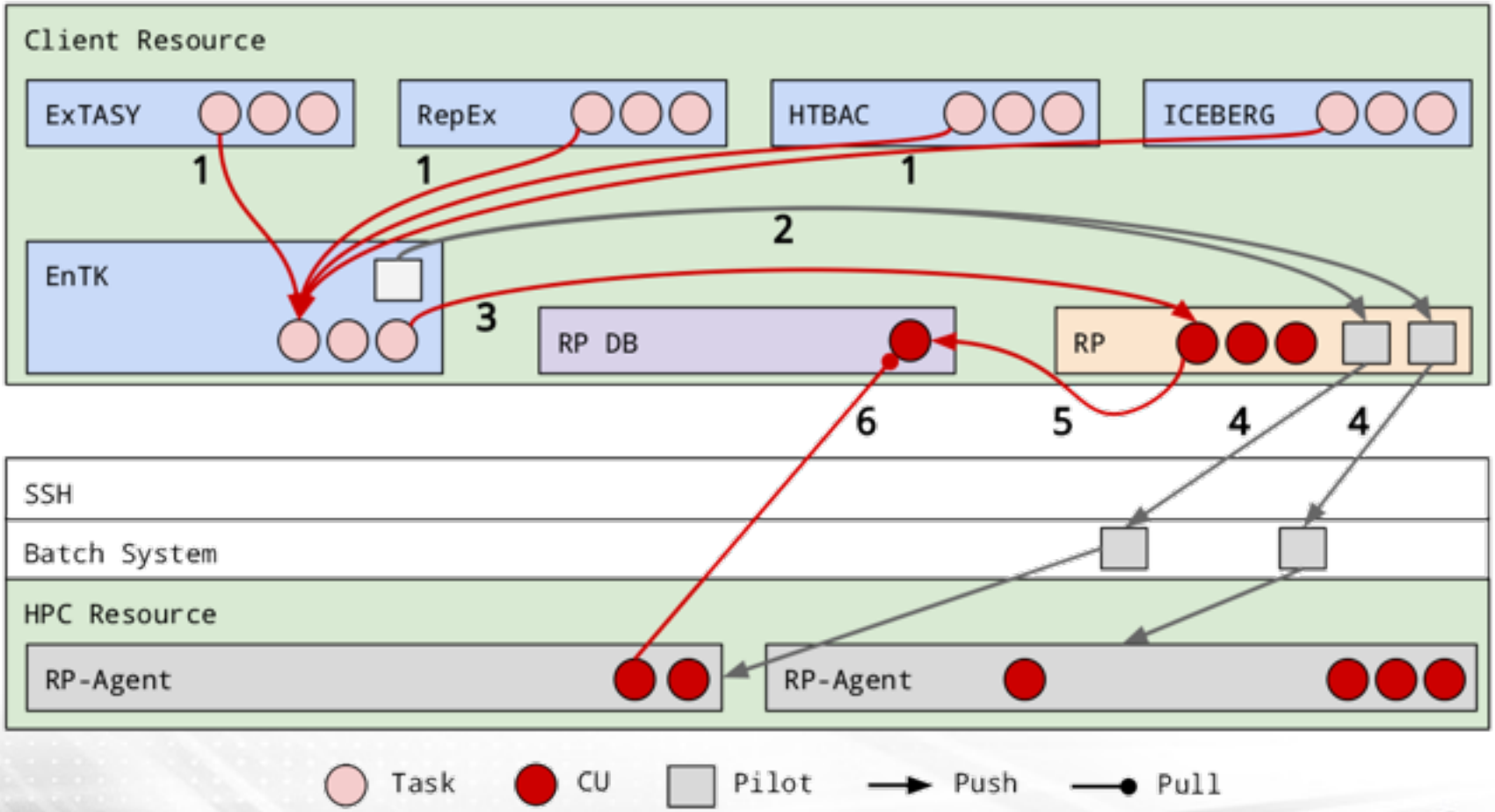
HTBAC: Execution Preparation via Ensemble Toolkit (EnTK)



(1) **TIES** (alchemical **protocol**) employs enhanced sampling at each lambda window to yield reproducible, accurate and precise relative binding affinities.

(2) **ESMACS** (endpoint **protocol**) is a computationally cheaper but less rigorous method used to directly compute the binding strength of a drug to the target protein from molecular dynamics simulations (as opposed to differences in affinity).

HTBAC: Managing the Ensemble Execution with EnTK and RP



Advantage of Adaptive Ensemble Algorithms

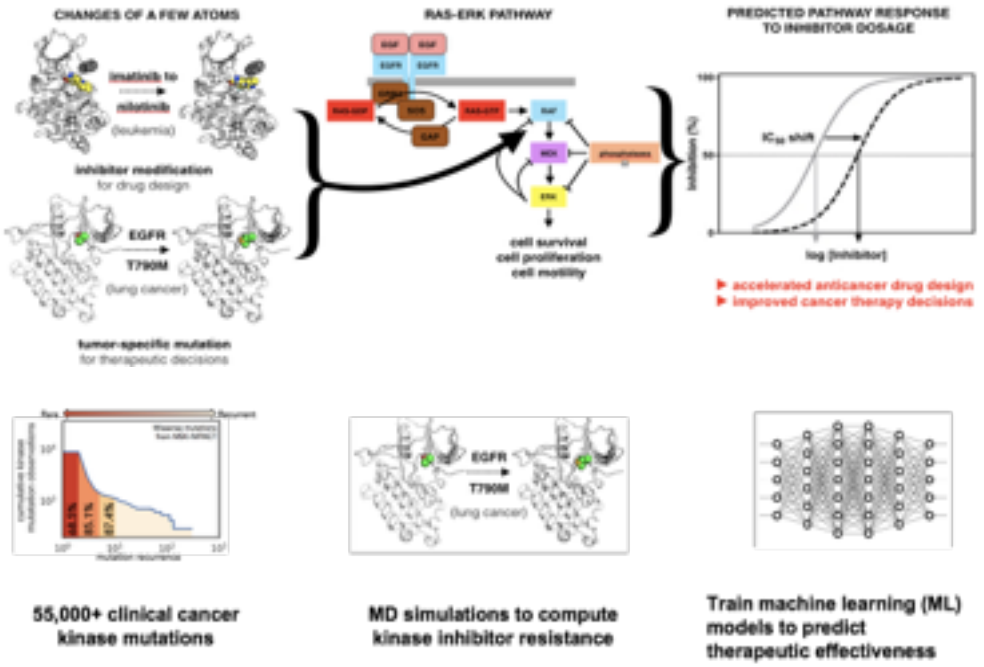
- Original study GSK and UCL
  - 16 drug candidates, BRD4 inhibitor
- Non-adaptive implementation of protocols on average:
  - ESMACS 10K core-hours
  - TIES 25K core-hours
  - O(1M) core hours without adaptivity
- **Using Adaptive Algorithms at scale resulted in 2.5x reduction in core hours needed**
- Computation performed on Titan and Blue Waters.

[1] S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, P. Bamborough, and P. V. Coveney. Rapid and reliable binding affinity prediction of bromodomain inhibitors: a computational study. *J. Chem. Theory Comput.*, 13(2):784–795, 2017.

[2] J. Dakka et al., Enabling Trade-offs Between Accuracy and Computational Cost: Adaptive Algorithms to Reduce Time to Clinical Insight, 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Washington, DC, USA, 2018, pp. 572-577. doi:10.1109/CCGRID.2018.00005

INSPIRE: Integrated (ML-MD) Scalable Prediction of REsistance

- Chemical space of drug design in response to mutations is large, 10K-100K mutations; too large for HPC simulations alone.
- Develop methods that use: 1) simulations to train machine learning (ML) models to predict therapeutic effectiveness and 2) ML models to determine which drug candidates to simulate.



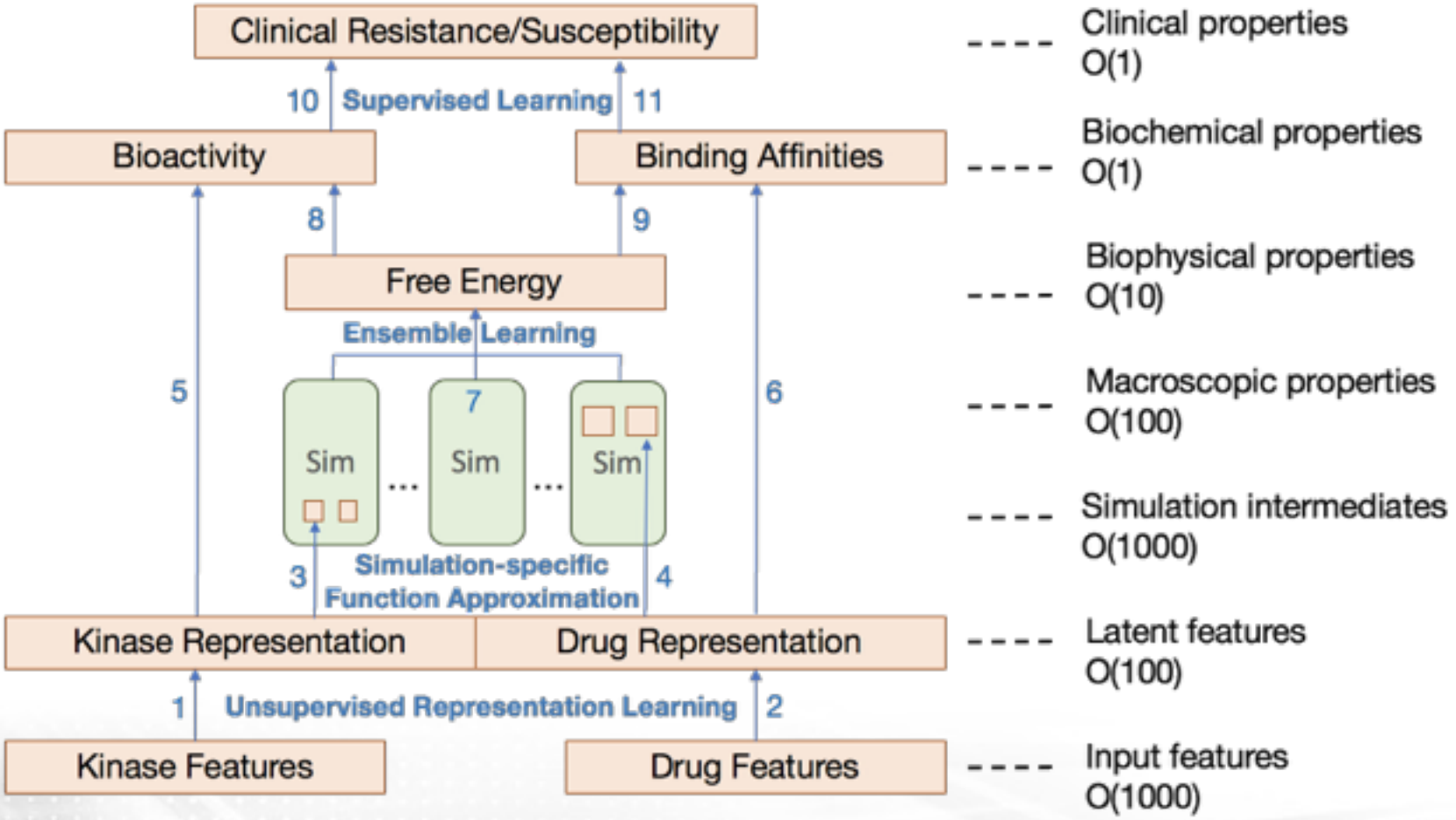
Early Science Project on NSF Frontera. DD Award on Summit.
A collaboration between BNL/Rutgers (Jha), Chicago (Stevens), Memorial Sloan Kettering (Chodera), UCL (Coveney)

Machine Learning @ Multiple Levels and Points in the Discovery Process

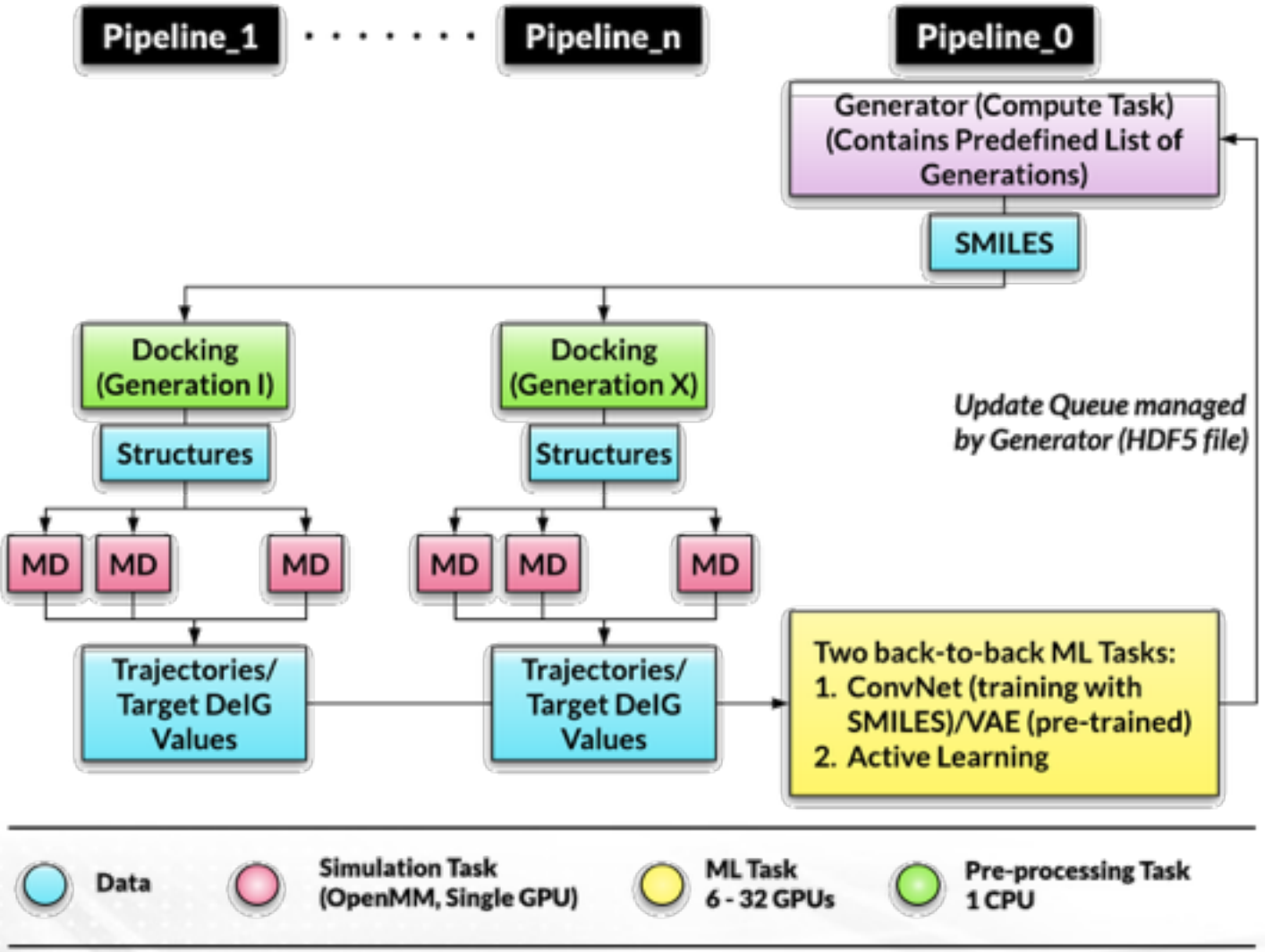
- Can ML enhance the **effective performance** of HPC simulations?
- Argue ML can enhance HPC simulations by 10^6 (?), if not greater.
- Enhancement not measured by Flops or usual performance measures, but **science done using same amount of computing for given accuracy**.
- Challenges:
  - System and application software
  - Application architecture and formulation

Learning Everywhere!
https://arxiv.org/abs/1902.10810

Machine Learning @ Multiple Levels and Points in the Discovery Process



INSPIRE


Pipeline_1
Pipeline_n
Pipeline_0
Generator (Compute Task)
(Contains Predefined List of Generations)
SMILES
Docking (Generation I)
Docking (Generation X)
Structures
Structures
MD
MD
MD
MD
MD
MD
Trajectories/ Target DeIG Values
Trajectories/ Target DeIG Values
Update Queue managed by Generator (HDF5 file)
Two back-to-back ML Tasks:
1. ConvNet (training with SMILES)/VAE (pre-trained)
2. Active Learning
Data
Simulation Task (OpenMM, Single GPU)
ML Task 6 - 32 GPUs
Pre-processing Task 1 CPU

RADICAL: Pilot on Leadership Class Machine

- Can we get performance agnostic of batch queue systems and MPI flavor?
  - LSF, PBS, SLURM, … ?
  - MVAPICH, … MPI flavors?
- **PMI-X**: **P**rocess **M**anagement **I**nterface for E**x**ascale, https://github.com/pmix/pmix/wiki
  - **PRRTE**: **P**MI-X **R**eference **R**un**T**ime **E**nvironment, https://github.com/pmix/prrte
- PMI used by MPI implementations, batch system
- Private DVM, concurrent tasks
- Pros: heterogeneous tasks (as with JSRUN); (potentially) fast, **portable**
- Cons: Young code; emerging official support

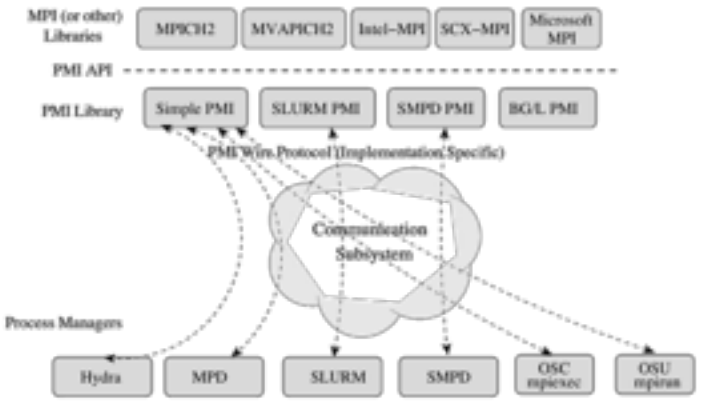
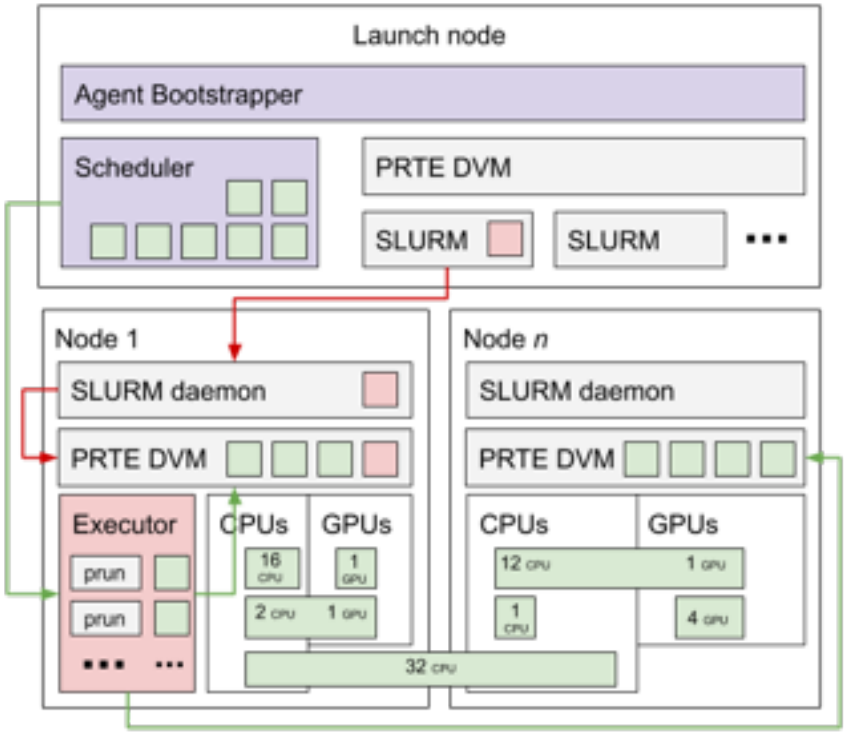


Fig. 1. Interaction of MPI and the process manager through PMI

Balaji P. et al. (2010) PMI: A Scalable Parallel Process-Management Interface for Extreme-Scale Systems. In: Keller R., Gabriel E., Resch M., Dongarra J. (eds) Recent Advances in the Message Passing Interface. EuroMPI 2010. Lecture Notes in Computer Science, vol 6305. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-15646-5_4

RADICAL: Pilot on Leadership Class Machine

- Can we get performance agnostic of batch queue systems and MPI flavor?
  - LSF, PBS, SLURM, … ?
  - MVAPICH, … MPI flavors?
- **PMI-X**: **P**rocess **M**anagement **I**nterface for E**x**ascale, https://github.com/pmix/pmix/wiki
  - **PRRTE**: **P**MI-X **R**eference **R**un**T**ime **E**nvironment, https://github.com/pmix/prrte
- PMI used by MPI implementations, batch system
- Private DVM, concurrent tasks
- Pros: heterogeneous tasks (as with JSRUN); (potentially) fast, **portable**
- Cons: Young code; emerging official support



Summary

- Brookhaven Lab is a leading DOE research laboratory with a focus on foundational physical, chemical and biological research in the DOE mission space.
- Brookhaven Lab has a history of pivoting research to create powerful innovations for the medical sector.
- Novel machine learning, artificial intelligence and high-performance computing research also are repurposed to support major advances in medical diagnostics and treatment, as well as optimized drug design.