Impacting Precision Medicine with DOE HPC, ML and AI Research

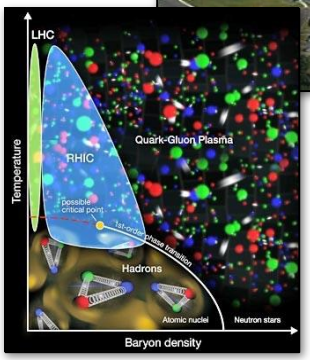
Nick D'Imperio
Brookhaven National Laboratory
Computational Science Initiative

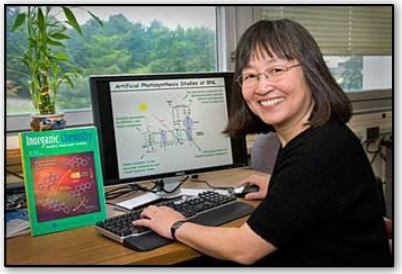
BROOKHAVEN NATIONAL LABORATORY

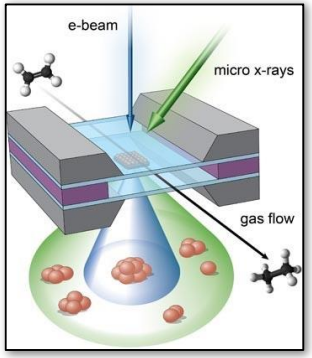
U.S. DEPARTMENT OF ENERGY

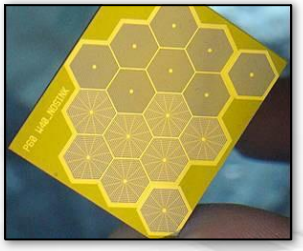
Brookhaven Lab at a Glance…

- Est. 1947
- One of 17 DOE Labs, only one in Northeast
- 2,600 employees => ~5000 jobs in NY State
  - 400 Grad/Undergrad Students
- $580M annual budget
- 5,322 acres with 316 buildings
- Over 2,600 facility users and 2,100 visiting scientists per year
- Fundamental, basic research to innovation, development and commercialization of technologies: energy S&T, nuclear and high energy physics, bio and environmental sciences, big data, national security









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 BNL study of relationship between trace elements and neurological diseases.
- Thallium-201, a radioisotope developed at BNL is used in **heart stress** tests world wide.
- 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.

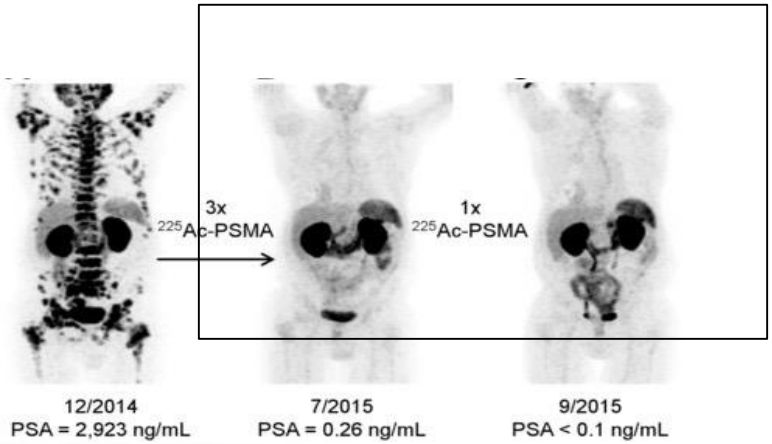
Nora Volkow is the former BNL Associate Director for Life Sciences who spent over 15 years at Brookhaven using PET technology to investigate the physical causes of addiction in the human brain.

Utilizing today's Research for Medical Progress

Brookhaven Linac Isotope Producer (BLIP)

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

BLIP



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

Pivoting BNL Machine Learning, Artificial Intelligence and High Performance Computing Research towards Precision Medicine

BNL operates many data-rich Experimental Facilities

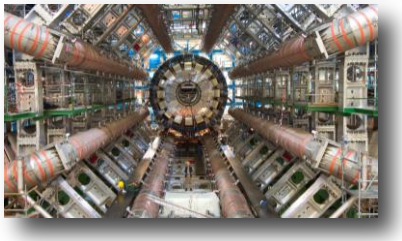
- Relativistic Heavy Ion Collider (**RHIC**)
- National Synchrotron Light Source II (**NSLS-II**)
- Center for Functional Nanomaterials (**CFN**)
- Accelerator Test Facility (**ATF**)
- Large Hadron Collider (LHC) **ATLAS**
- Atmospheric Radiation Measurement (**ARM**) program
- **Belle II**: computing for neutrino experiment
- Quantum chromodynamics (**QCD**) computing facilities for BNL, RIKEN, & US QCD communities

RHIC

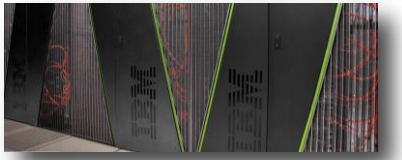
NSLS II

CFN

ATLAS

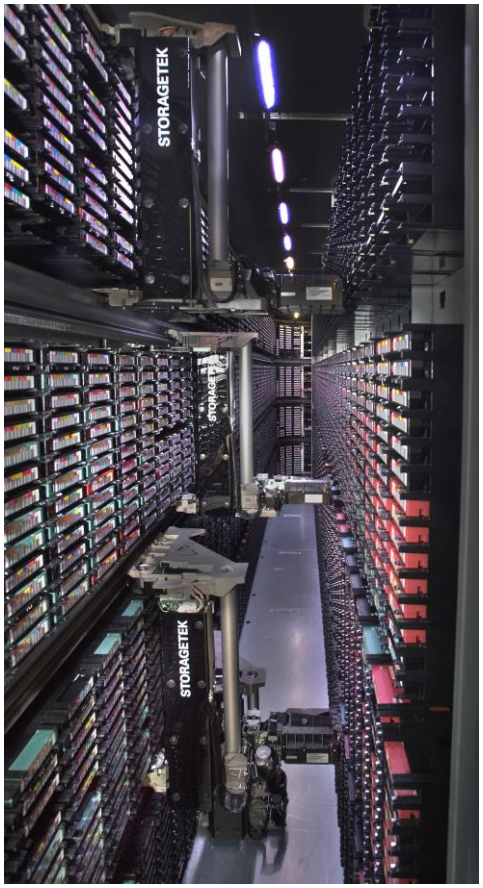


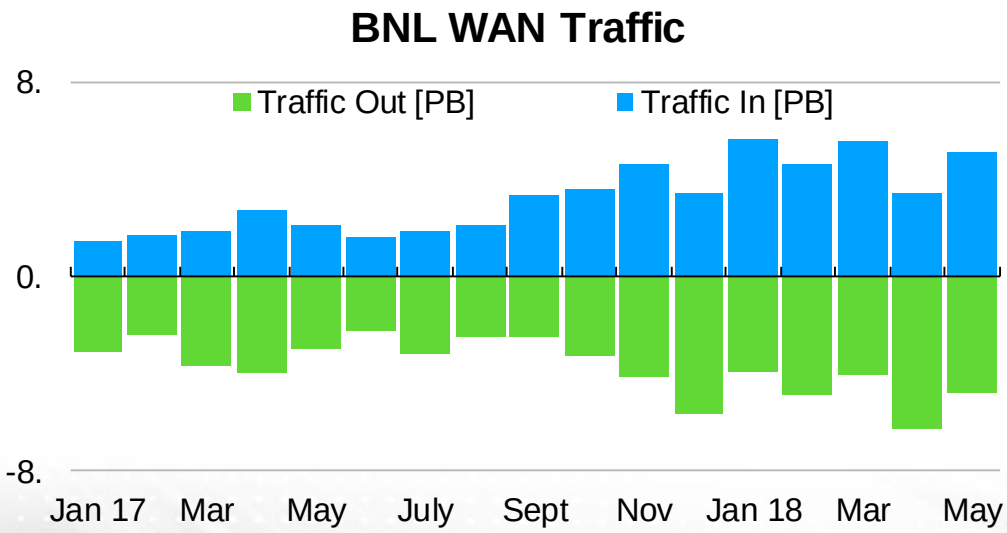
QCD



BNL Data in Numbers …

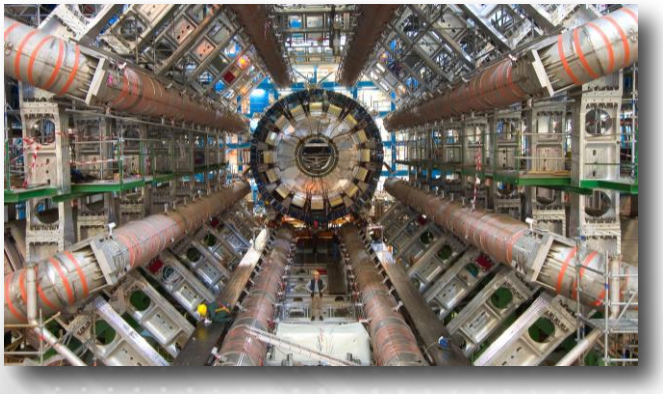
- **Largest scientific archive in the US**, third largest in the world (ECMWF, UKMO)
- **~145 PB archived** to date
- **~110PB Data Transfer per Year**





Key Challenge Examples at our Large Scale Facilities

- Real Time Analysis and Steering of Experiments – Challenges:
  - CFN - 400 images/sec
  - NSLS II – up to 5TB/s in burst



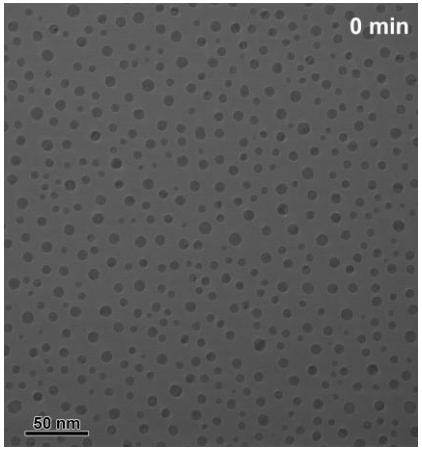
- Extreme Scale Data Management and Analysis:
  - **690 PB of data analyzed in FY18**
  - Moved data to compute and storage in 2016 – 1.6 Exabyte

Machine Learning and Artificial Intelligence are Key to Making Sense of Data @ Scale

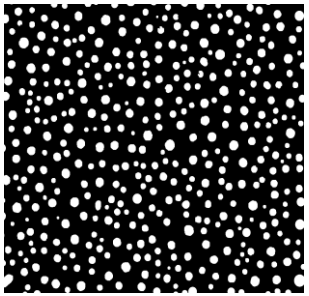
- Research Leaders in real time, unsupervised, robust machine learning methods

- Many new ML methods developed for real time analysis and interpretation of experimental data / images to help direct experiments as they are developing.

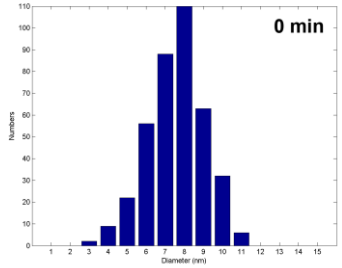
- Additional methods created that can integrate and interpret data from multiple, different sources.



Original images



Detection results



Size distribution

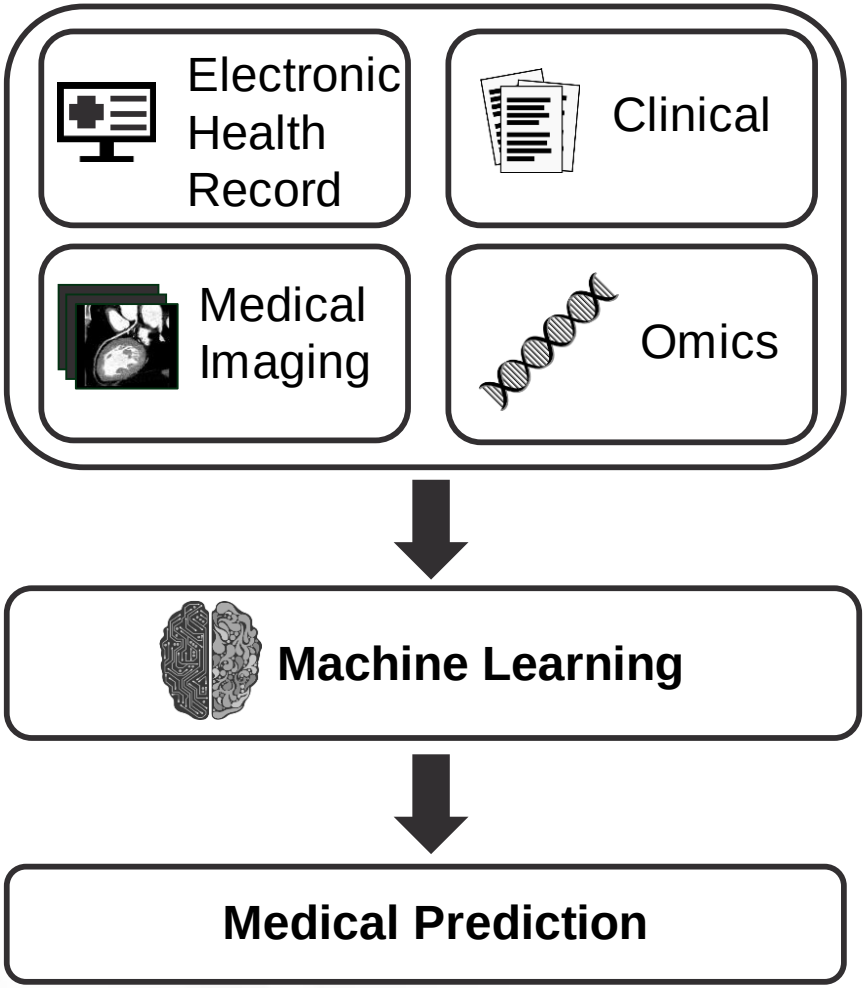
CFN - Nanoparticle Detection in TEM Images

Machine Learning for Medicine

NHIS Ilsan Hospital
- Dementia
- Depression

Veteran Affairs
- Prostate Cancer
- Cardio Vascular Disease
- Suicide Prevention

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



Collaboration: DOE – Veteran's Affairs (VA)

- Provide:
  - Scalable Algorithms
  - Integration of large heterogenous Data Sets
    - Images
    - Full Genome
    - Health Records
    - Longitudinal Studies
  - Uncertainty Quantification and Error Analysis
- **Goal: Applying DOE developed HPC, Simulation and large scale Data Analysis Capabilities to VA data to improve healthcare for our Veterans**

DOE – VA - Targets

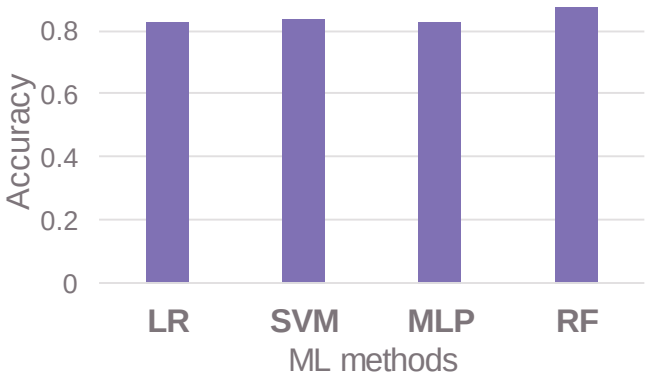
- ***Enhanced prediction and diagnosis of Cardiovascular Disease (CVD)***
  - Develop methods to inform individualized drug therapies to prevent, pre-empt and treat CVD.
  - Enhance prediction, diagnosis and management of major CVD subtypes in Veterans
- ***Precision discrimination of lethal from non-lethal Prostate Cancer***
  - Build improved classifiers to distinguish lethal from non-lethal prostate cancers.
  - Reduce unnecessary treatments /provide an increased quality of life for patients
- ***Patient-specific analysis for Suicide Prevention***
  - Provide tailored and dynamic suicide risk scores for each Veteran at risk.
  - Create a clinical decision support system that assists VA clinicians in suicide prevention efforts, and helps to evaluate effectiveness of various prevention strategies

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 by utilizing only Electronic Health Records data, and the approach can be applied to other types of cancer too.

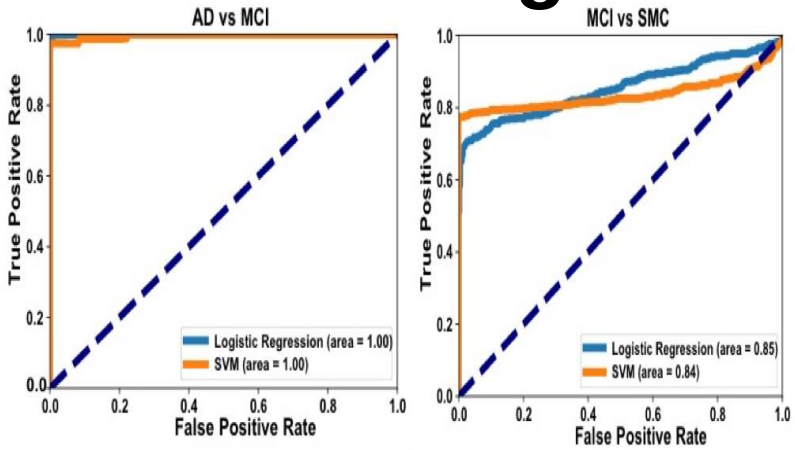
Exploited patients' demographics, disease and medication histories, and physical examination records to predict prostate cancer



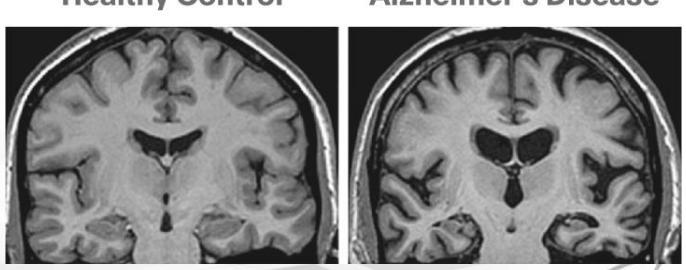
Comparison of machine learning algorithms for predicting prostate cancer occurrence in 1 or 2 years. We used four algorithms in the comparison: logistic regression (LR), support vector machine (SVM), multilayer perceptron (MLP), and random forest (RF). We achieved high prediction accuracy in all algorithms without using other types of data. The result showed that our task is easily verifiable and robust.

Accurate Prediction of Alzheimer's Disease with novel Machine Learning

- Brookhaven National Laboratory (BNL) 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
- With the algorithms, BNL and its collaborators—Columbia University and the National Health Insurance Service (NHIS) Ilsan Hospital in South Korea—**achieved nearly 100% Alzheimer's detection accuracy and 83% prediction accuracy for early-stage Alzheimer's**



Receiver-Operator Characteristic (ROC curve) Analysis. Combined models (morphometry + connectome) accurately classified AD (Alzheimer's Disease), MCI (Mild Cognitive Impairment), and SMC (Subjective Memory Complaints). (a) AD vs SMC classification; (b) MCI vs SMC classification. Logistic Regression and SVM classifiers have similar area-under-curve (AUC) for different binary classifications.



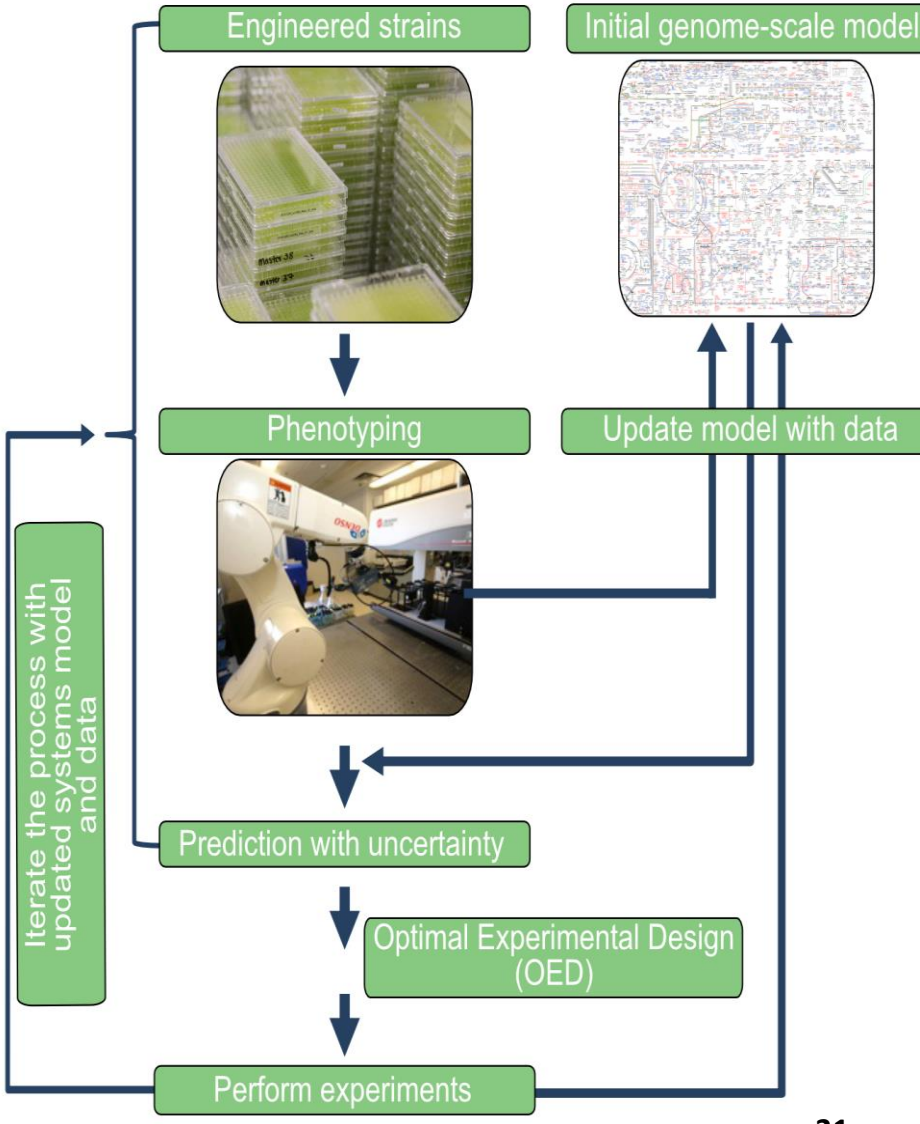
Artificial Intelligence-driven Optimal Experimental Design (OED)

Example problems

- Improve biofuel yield for algal systems
- Determine phase diagrams for complex materials
- Use HPC resources efficiently for costly biomolecular simulations
- Optimize RHIC Beam Energy Scan data taking strategy
- Drug Discovery and Optimization

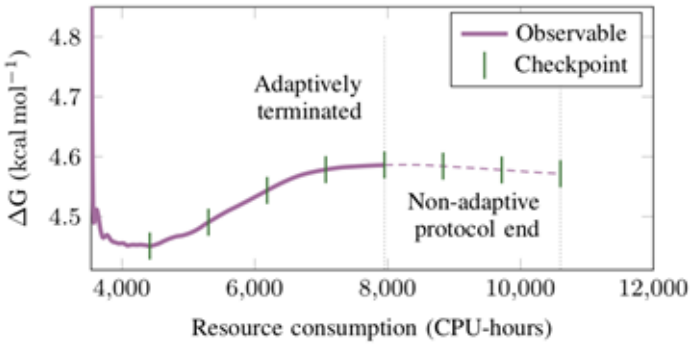
OED for Biofuel Yield Optimization

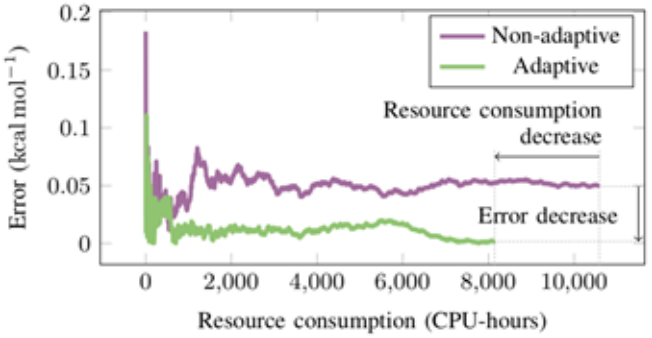
- It is not feasible to experimentally characterize every sequenced gene
- In bioenergy crops, most functional annotations are based on sequence similarity alone
- Most protein function characterizations are from studies with bacterial, yeast and animal genes/proteins
- Often, data used to inform gene/protein function is qualitative or vague, limiting its veracity and value when propagated to other organisms



OED for Drug Design

- Binding free energy is measure of goodness of compounds binding with their target proteins; computed using ensembles of Molecular Dynamics (MD) simulations
- High-Throughput Binding Affinity Calculator (HTBAC): Adaptive workflow systems to terminate & substitute simulations
- Significant improvement in "efficiency" on test data by GSK (in collab. with UCL)
- **IEEE SCALE2018 Award**

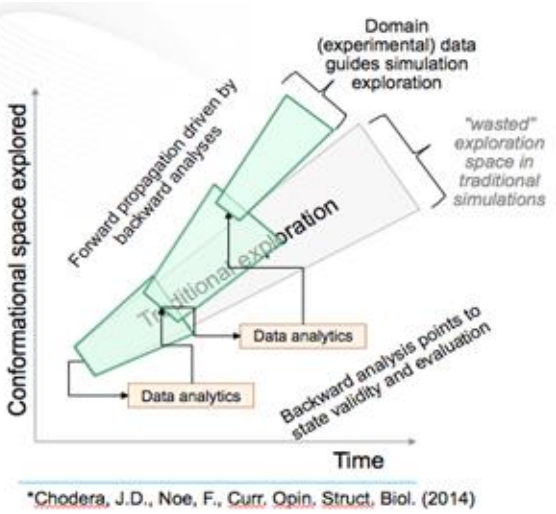




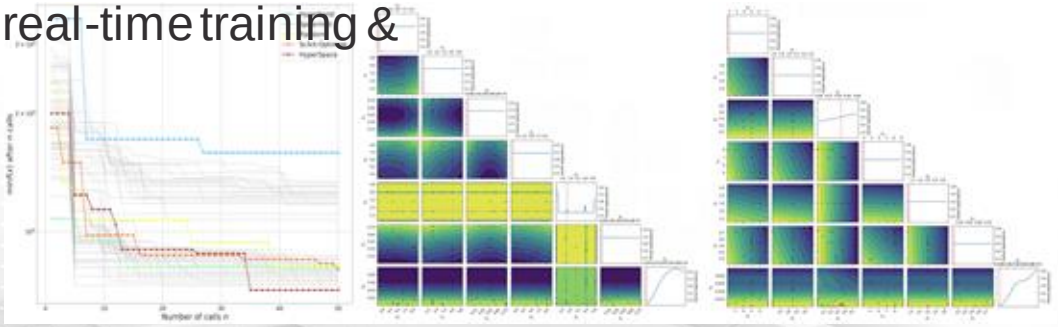
- Ongoing steps: OED Drug Design using Reinforcement Learning:
- Use Bayesian Learning to determine algorithms & parameters adaptively

Using Deep Learning to improve MD Simulations

- National Cancer Institute (NCI) -DOE Pilot
- ML-Driven (Classical) Molecular Dynamics (MD)
  - Schematic shows how active learning improves sampling of MD simulations
    - Deep clustering of protein folding simulations using CVAE (ORNL) and Bayesian Hyperparameter Optimization using RADICAL-Learn on Summit Building low dimensional representations of states from simulation trajectories.
    - CVAE can transfer learned features to reveal novel states across simulations
- DL approaches to achieve near real-time training & prediction!



Deep clustering of protein folding simulations, Debsindhu Bhowmik et al, https://doi.org/10.1101/339879



Summary

- BNL is a leading DOE research laboratory, with a focus on foundational physical, chemical and biological research in the DOE mission space.

- We have a history of pivoting some of our research to created powerful innovations for the medical sector.

- Novel Machine Learning, Artificial Intelligence and High performance Computing Research is now also repurposed to support major advances in medical diagnostics and treatment, as well as optimized drug design.