Changing Science through Online Analysis

Kerstin Kleese van Dam, Computational Science Initiative Director

Huub van Dam, Shinjae Yoo, Shantenu Jha, Wei Xu, Sangsoo Ha

BROOKHAVEN NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

Breaking the Paradigm

- Until recently subscribed to the ***post-hoc*** analysis paradigm
- This meant that Calculations incl. Workflows:
  - Had to be fully planned out at conception
  - Could only adapt to pre-conceived options
  - Were only evaluated in terms of success, scientific outcome etc. after they had completed
- However, **scientific discovery relies on the rare, the unexpected!**

Online analysis can more effectively support scientific discovery

Online Adaptive Detection of Events

BROOKHAVEN
NATIONAL LABORATORY

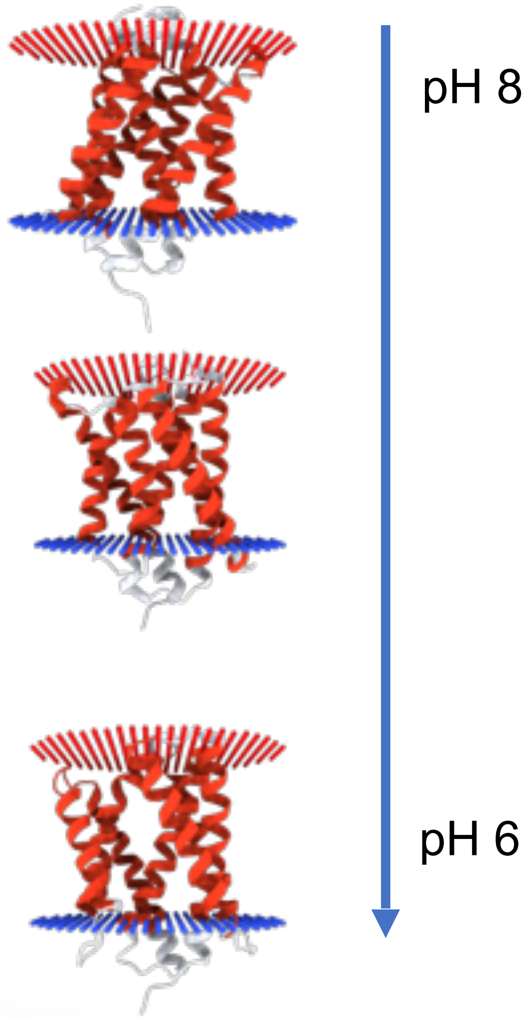
U.S. DEPARTMENT OF
ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

NWChemEX

Initial Use Case - Addressing Data Challenge at Exascale

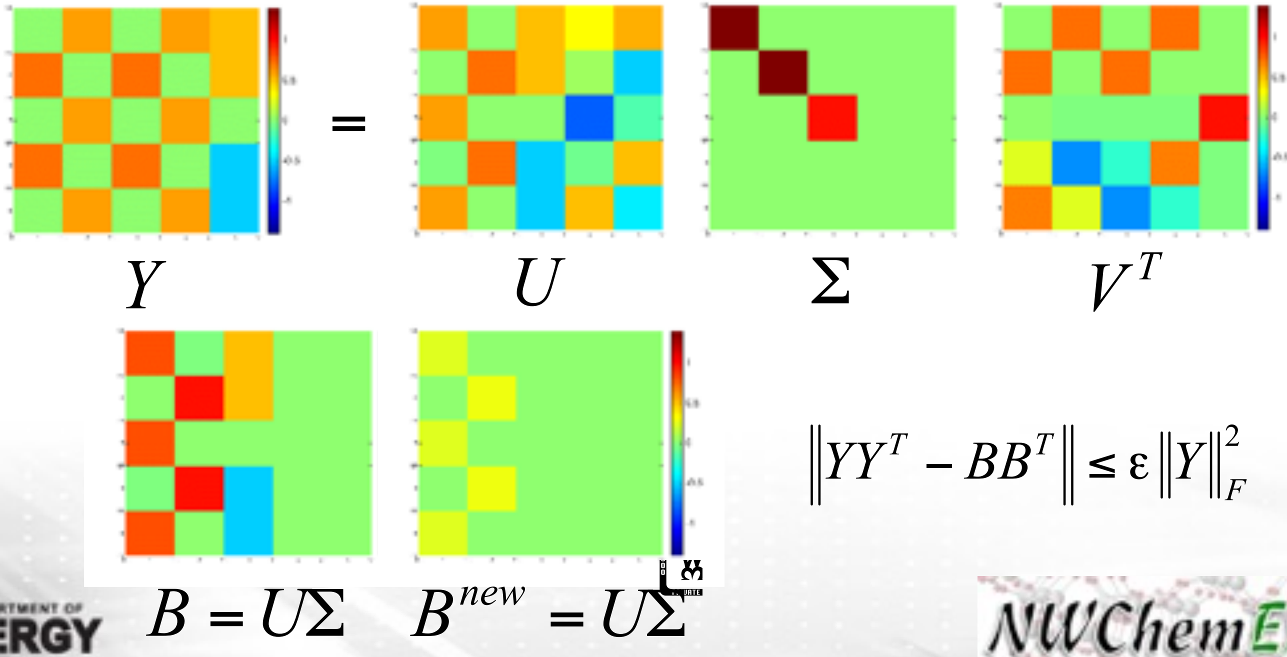
- 1 Million atom simulations
- Conformational changes take about a microsecond (1e^{-6})
- Time resolution of the simulation is a femtosecond (1e^{-15})
- Runtime ~12 days
- Storing everything 8*4*1e^{6}*(1e^{-6}/1e^{-15})=**32 PB**
- Typical approach: uniform sampling
  Store 1 out of 1000 structures
- **Store too much and data volumes become overwhelming, store too little and you might miss the important transformations**



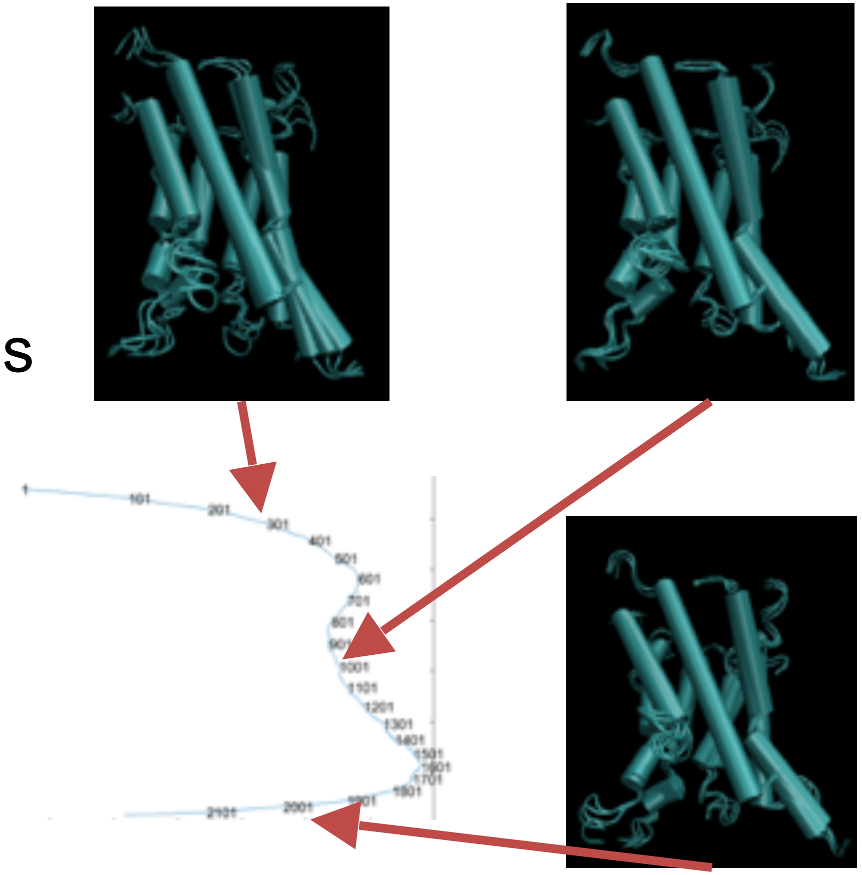
Y. Chang, R. Bruni, B. Kloss, Z. Assur, E. Kloppmann, B. Rost, W. A. Hendrickson, Q. Liu. *Structural basis for a pH-sensitive calcium leak across membranes.* Science **344**, pp. 1131-1135.

Online Adaptive Approach

- Manifold Learning for MD trajectories
  - Matrix Sketch of MD trajectories
- Important / Interesting event detection
  - Weighted Reservoir Sampling of gradient changes



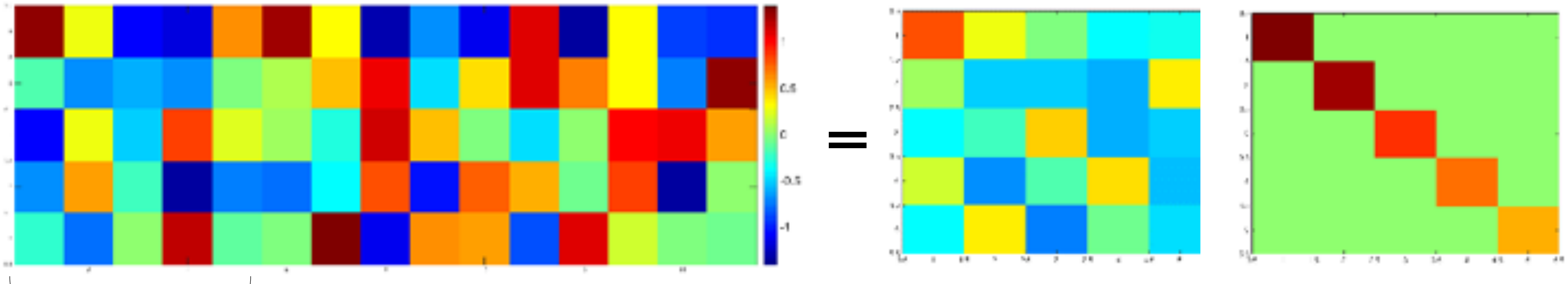
$$\left\| YY^T - BB^T \right\| \leq \varepsilon \left\| Y \right\|_F^2$$



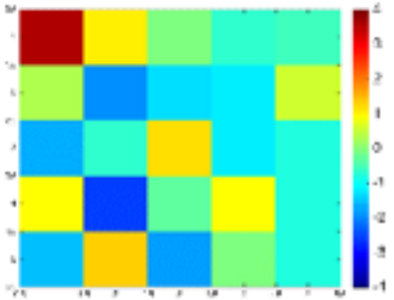
Low dimensional manifold projection of different state of MD trajectories

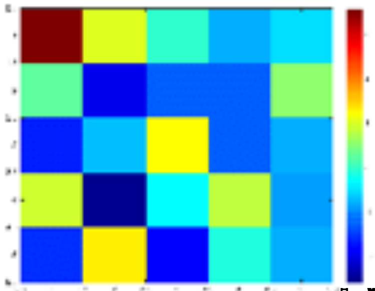
Streaming Single Value Decomposition

- Our Extension - Why only sampling one stream at a time?



$U_{(l)}$ $\Sigma_{(l)}$





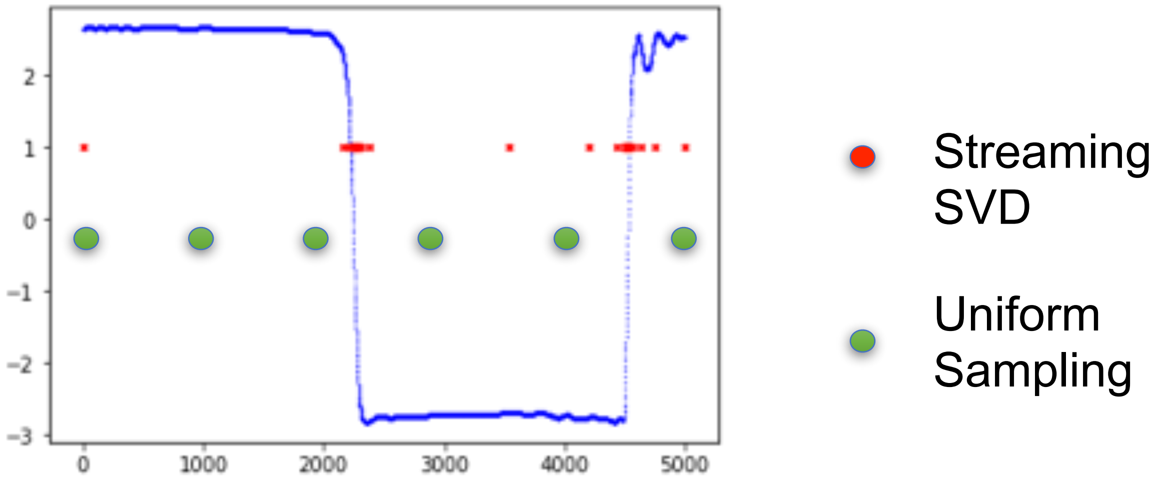
$B = U_{(l)}\Sigma_{(l)}$ $B^{new} = U_{(l)}\Sigma_{(l)}$

$$\ell = \Omega\left(\frac{\sqrt{m}\|Y_{[\ell]}\|^2\|Y_{[\ell]} - Y_{[\ell]_{(k)}}\|_F^2}{L^2}\right)$$

$$\|\mathbf{w}_\ell - \tilde{\mathbf{w}}_\ell\| \leq \frac{k\|Y_{[\ell]} - Y_{[\ell]_{(k)}}\|_F}{\breve{\sigma}_{\ell_k}^2 + \alpha}\sqrt{\frac{\Gamma_a\Gamma_b}{\ell - k}} + \frac{\sigma_{t_k}}{\sigma_{\ell_k}^2 + \alpha}\frac{\sqrt{2}L}{\sqrt{L + 8\kappa^2\|Y_{[\ell]}\|^2\sqrt[4]{L^2 + 16\kappa^4\|Y_{[\ell]}\|^4}}}$$

A Case Study - 32 Samples

- Identifying important events when they occur, every time



NWChemEX - Achievement and Next Steps

- Orders of magnitude data reduction
- **Able to detect all points of interest - never miss an event!**
- Currently applied to classical MD simulations - science requires more accurate treatment of areas of interest
- Online analysis can enable us to fire off these methods when points of interest are detected, rather than after the end of a standard calculation, which would take ~12 days (for 1 microsecond of observation time)

Online Anomalie Detection

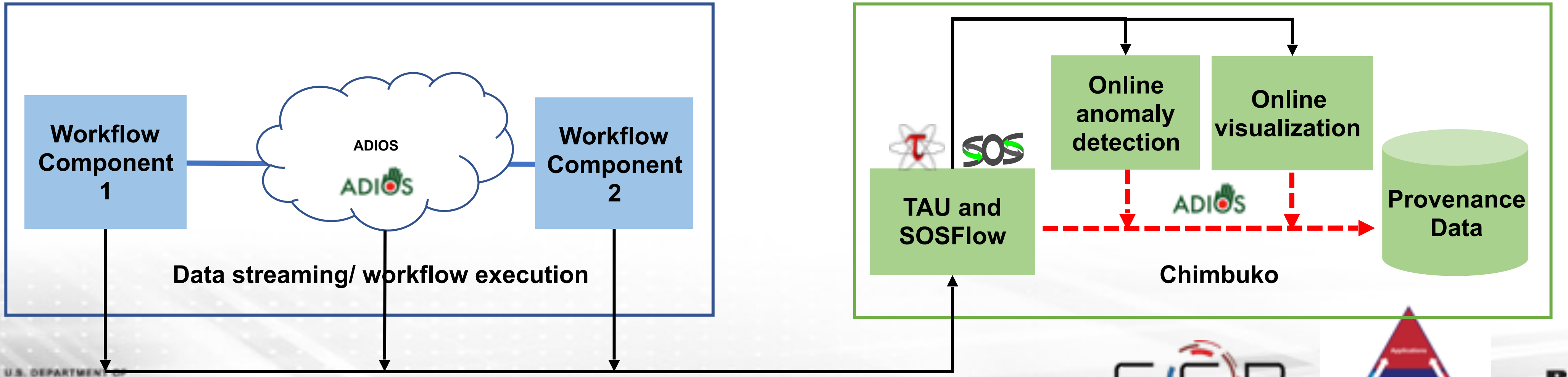
BROOKHAVEN
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

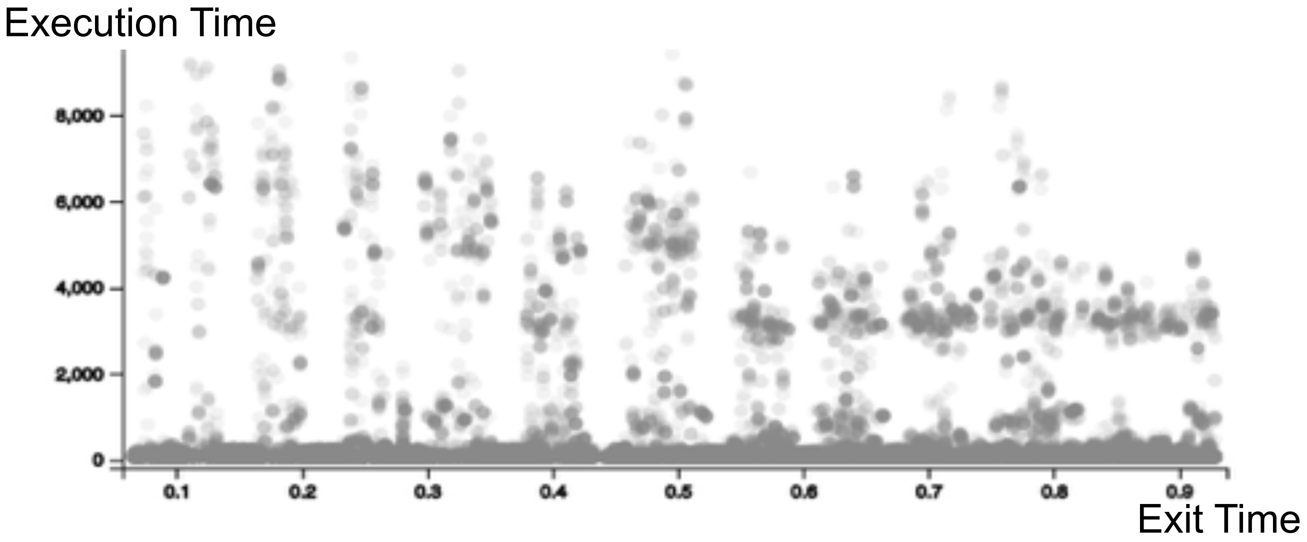
Chimbuko - Online PerformanceTrace Data Analysis at Exascale

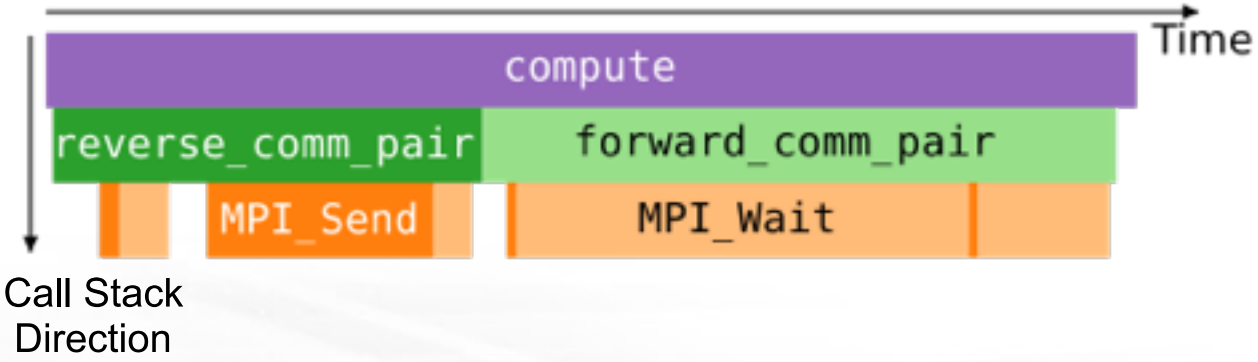
- Complex architectures, applications and workflows, require detailed performance analysis
- However, performance analysis at Exascale hits the same data challenges as the science apps.
- Our solution - online analysis and reduction to events of interest



Execution Time Based Anomaly Detection

- Execution time-based detection
  - Statistics approach (e.g., confidence interval)
  - Density-based (e.g., local outlier factor)
- Functions depend on each other
  - Delay of child functions
  - Communication delay of other nodes

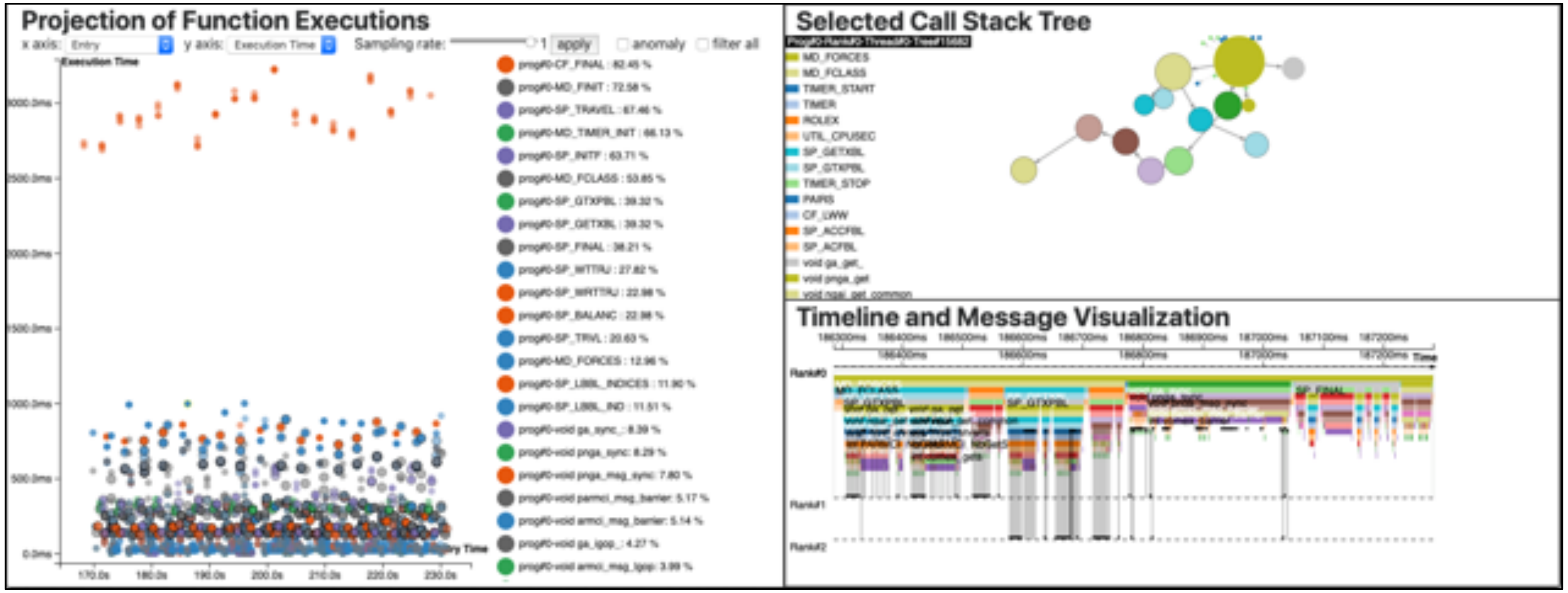




Streaming Performance Visualization

1. Streaming Workflow Overview in Scatter Plot

3. Function Execution in Dynamic Call Stack Tree

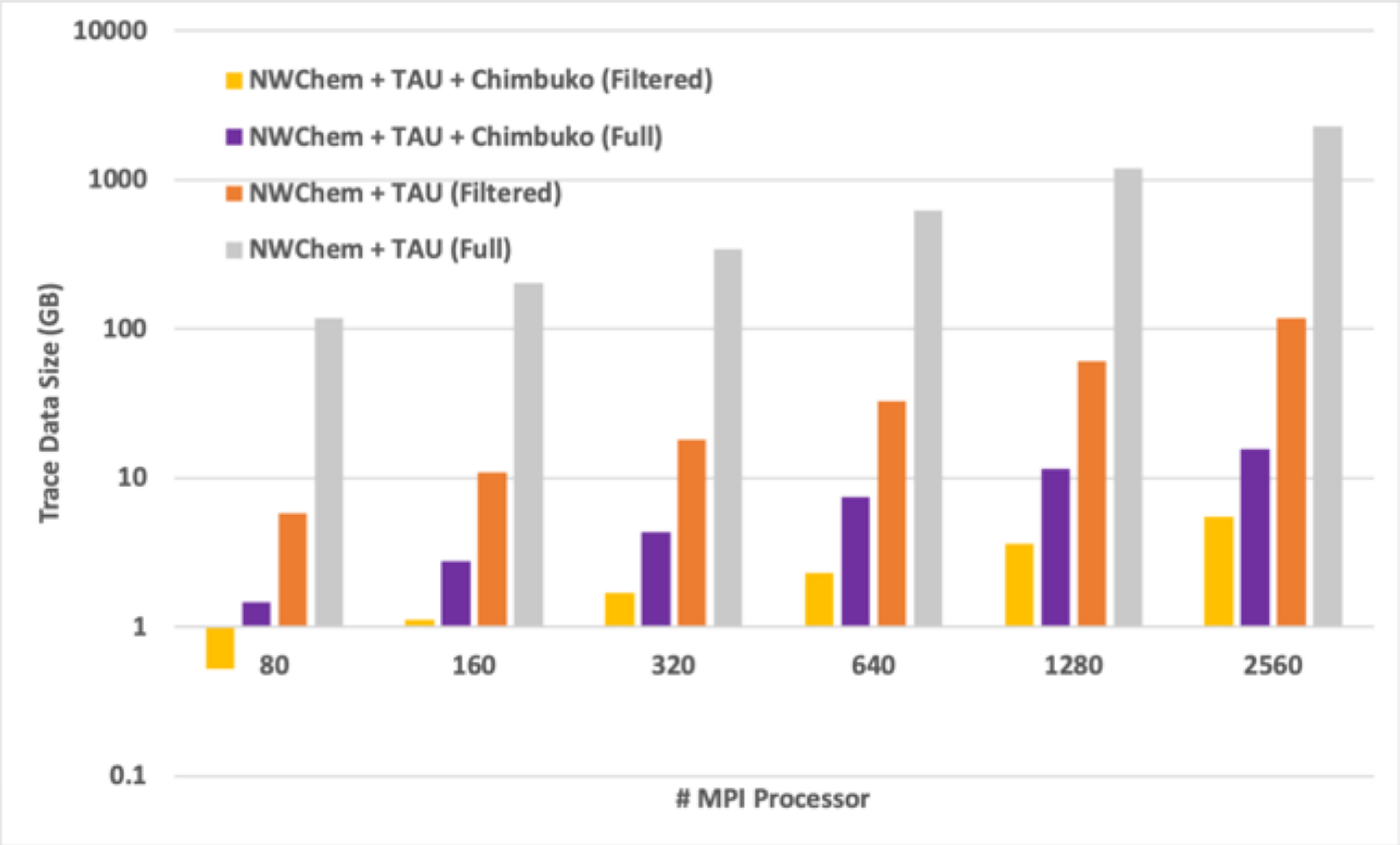

Projection of Function Executions
x axis: Entry
y axis: Execution Time
Sampling rate:
1
apply
□ anomaly
□ filter all
Execution Time
3000.0ms
2500.0ms
2000.0ms
1500.0ms
1000.0ms
500.0ms
0.0ms
Entry Time
170.0s
180.0s
190.0s
200.0s
210.0s
220.0s
230.0s
● prog#0-CF_FINAL : 82.45 %
● prog#0-MD_FINIT : 72.58 %
● prog#0-SP_TRAVEL : 67.46 %
● prog#0-MD_TIMER_INIT : 66.13 %
● prog#0-SP_INITF : 63.71 %
● prog#0-MD_FCLASS : 53.85 %
● prog#0-SP_GTXPBL : 39.32 %
● prog#0-SP_GETXBL : 39.32 %
● prog#0-SP_FINAL : 38.21 %
● prog#0-SP_WTTRU : 27.82 %
● prog#0-SP_WRTTRU : 22.98 %
● prog#0-SP_BALANC : 22.98 %
● prog#0-SP_TRVL : 20.63 %
● prog#0-MD_FORCES : 12.96 %
● prog#0-SP_LBBL_INDICES : 11.90 %
● prog#0-SP_LBBL_IND : 11.51 %
● prog#0-void ga_sync_ : 8.39 %
● prog#0-void pnga_sync: 8.29 %
● prog#0-void pnga_msg_sync: 7.80 %
● prog#0-void parmci_msg_barrier: 5.17 %
● prog#0-void armci_msg_barrier: 5.14 %
● prog#0-void ga_igop_: 4.27 %
● prog#0-void armci_msg_igop: 3.99 %
Selected Call Stack Tree
MD_FORCES
MD_FCLASS
TIMER_START
TIMER
ROLEX
UTIL_CPUSEC
SP_GETXBL
SP_GTXPBL
TIMER_STOP
PAIRS
CF_LWW
SP_ACCFBL
SP_ACFBL
void ga_get_
void pnga_get
void ngai_get_common
Timeline and Message Visualization
186300ms
186400ms
186500ms
186600ms
186700ms
186800ms
186900ms
187000ms
187100ms
187200ms
Time
Rank#0
Rank#1
Rank#2
MD_FCLASS
SP_GTXPBL
SP_FINAL

2. Dynamic functions of interest

4. Function Execution and Message Passing in Zoomable Timeline

- Streaming data reduction and aggregation
- Sliding time window of workflow overview for regular and anomaly function executions

Chimbuko Data Reduction > 100x



Chimbuko - Achievements and Next Steps

- First Online Performance Trace Data Analysis tool, not only for single applications, but also full workflows
- Enables real time identification and analysis or performance anomalies in realistic calculations
- Lightweight version could also support adaptive resource management for complex workflows to ensure optimal resource usage and guarantee of execution time constraints.

In-situ Compression Artifact Removal

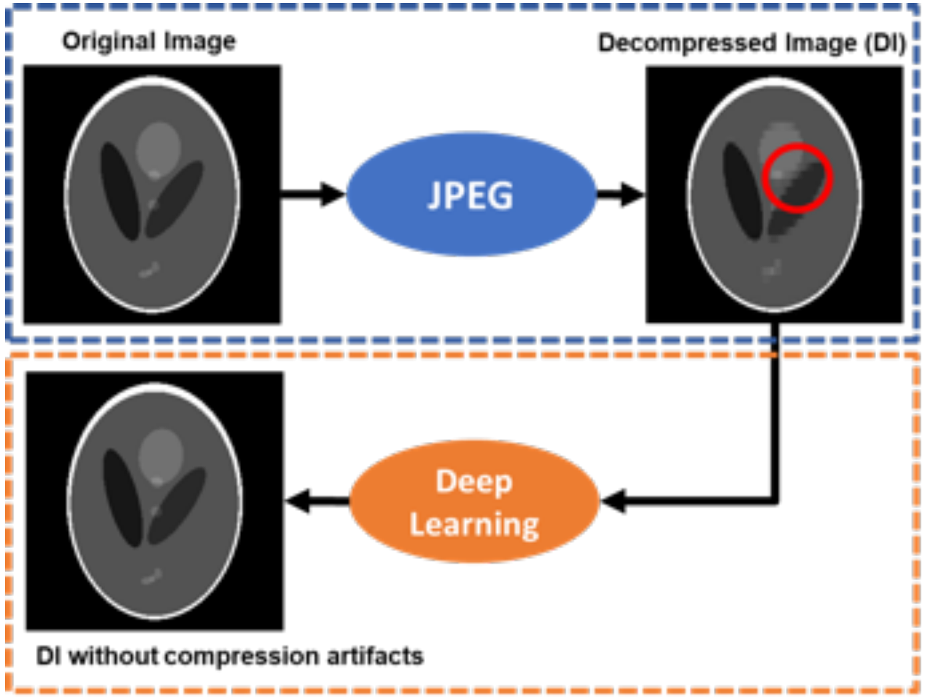
BROOKHAVEN NATIONAL LABORATORY

RAPIDS · Oak Ridge National Laboratory · Northwestern University · Argonne National Laboratory · U.S. Department of Energy

BROOKHAVEN SCIENCE ASSOCIATES

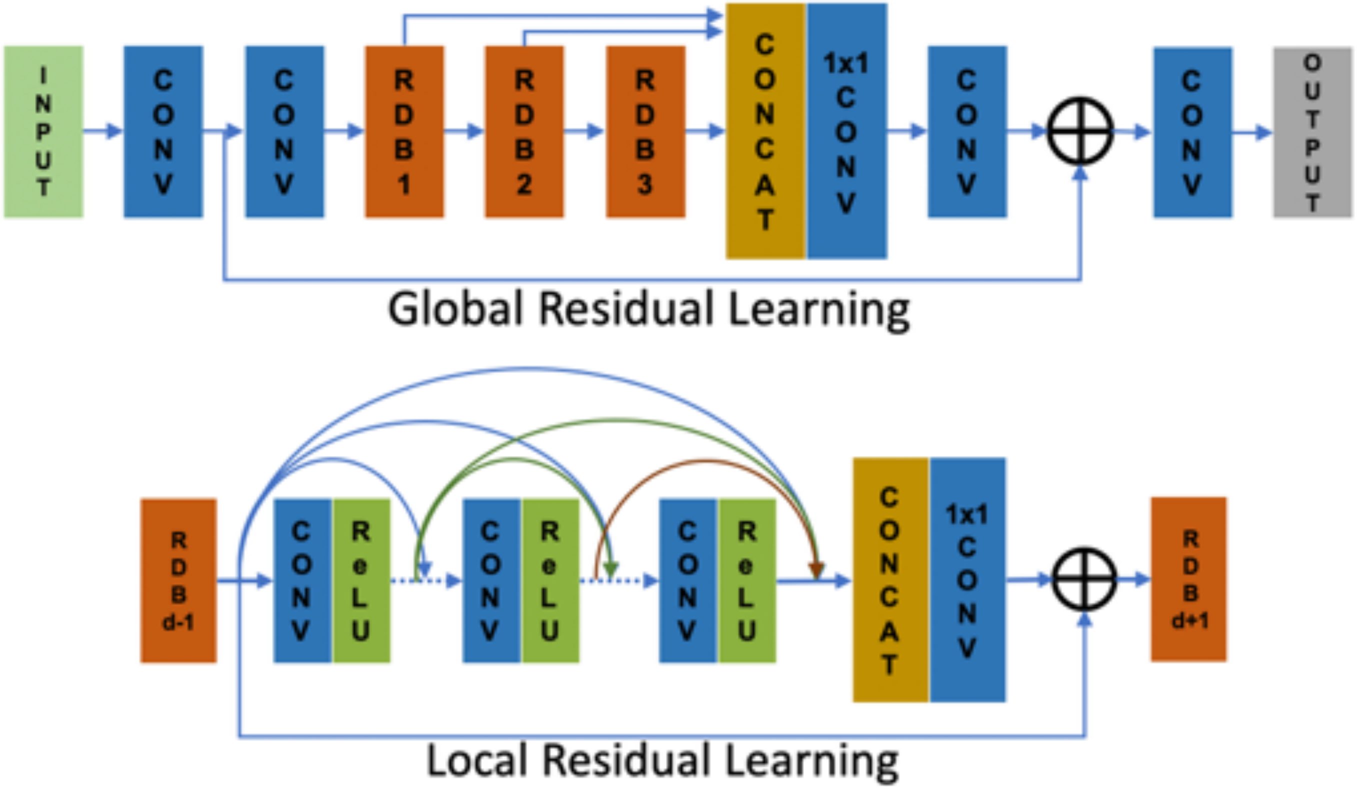
Why compression artifact removal?

- Many Exascale Applications will create vast volumes of **data, too big to store** (I/O constraints)
  - Scientific simulations (fusion, fluid dynamics, climate)
  - Experimental measurement devices (neutron / light sources)
- **Lossy image compression** is
  - Often used to achieve high data reduction rates
  - However, **leads to blocking artifacts and blurring**
- If we can **remove compression artifacts** in-situ, we can
  - **Boost scientific accuracy, while maintaining high compression rates**



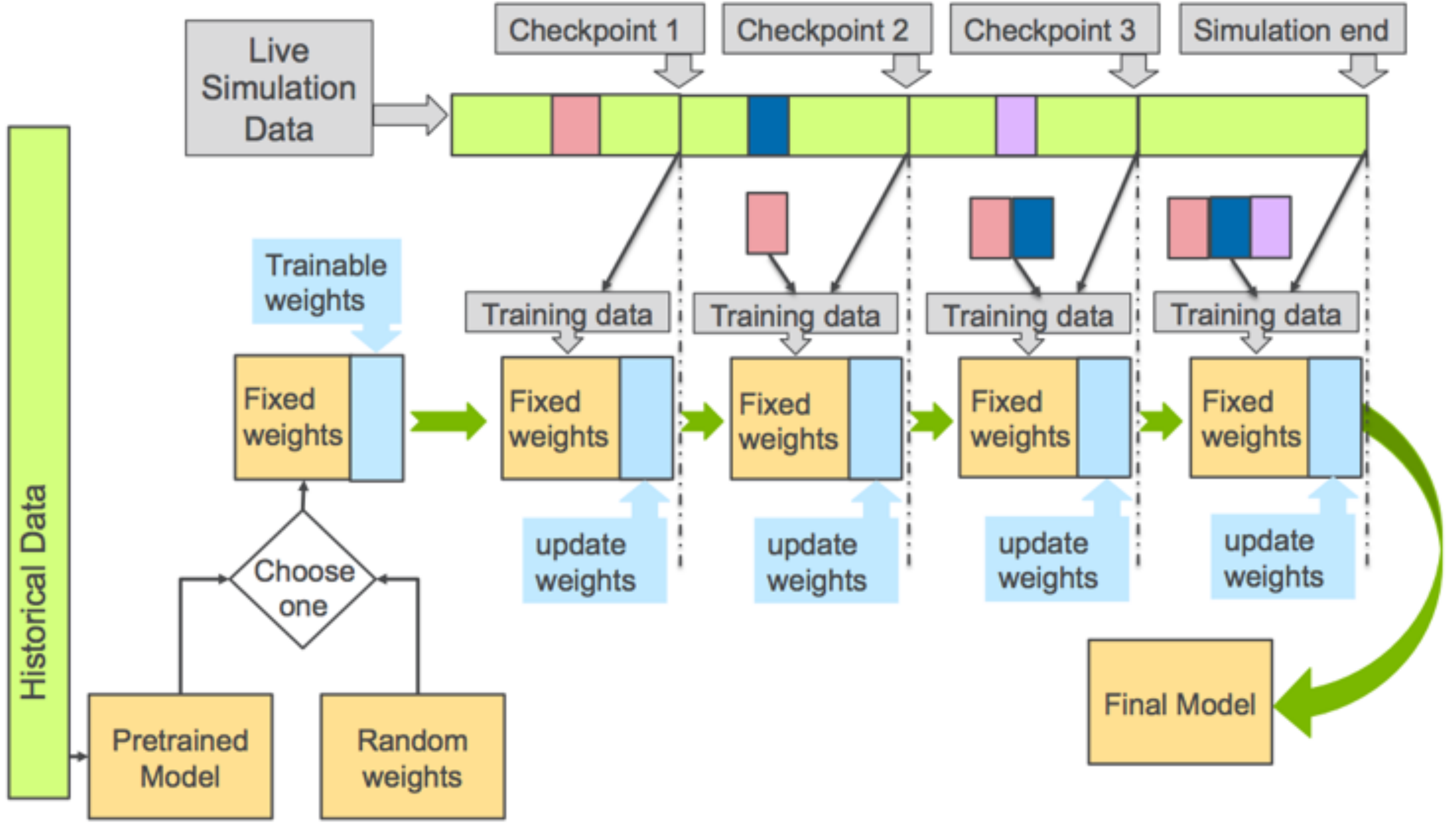
Residual Dense Net (RDN)

- **Speed:** Unlike a traditional approach (compressed sensing), DL approaches can reduce inference time significantly → Enable in-situ inference
- **Quality**: Residual Dense Net (RDN) shows better reconstruction than traditional approaches and other deep learning approaches (e.g., EDSR)



An overview of RDN model: global residual learning (top), and local residual learning between residual dense blocks (bottom)

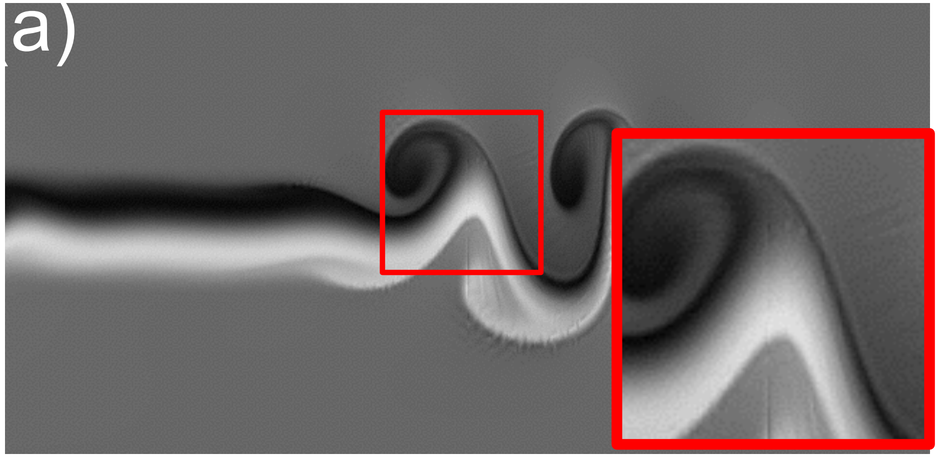
Incremental Batch Transfer Learning (IBTL)



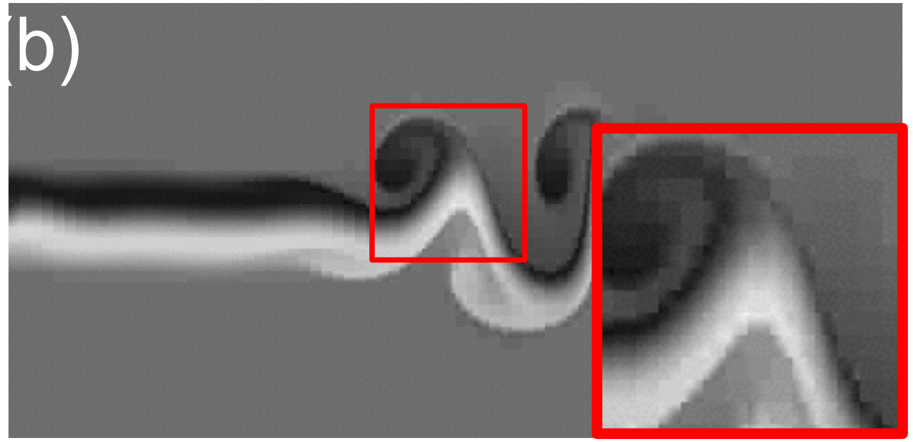
RAPIDS

BROOKHAVEN NATIONAL LABORATORY

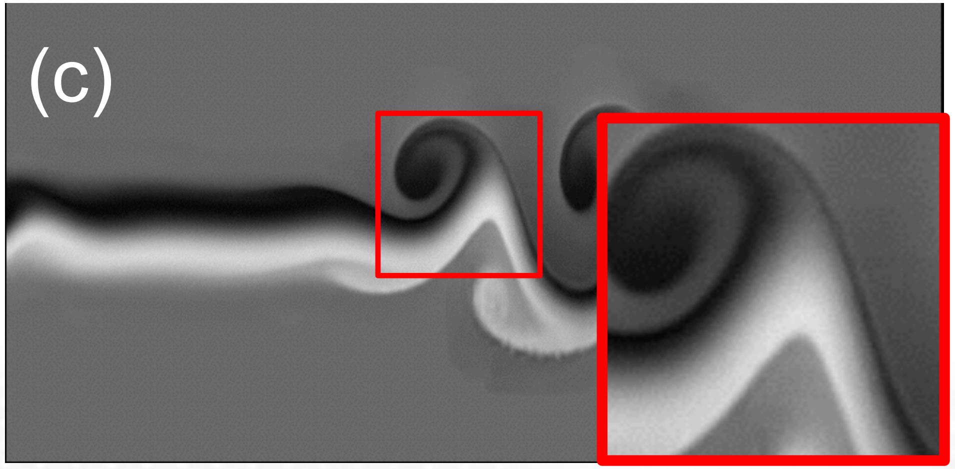
Case Study - Climate Data



Original

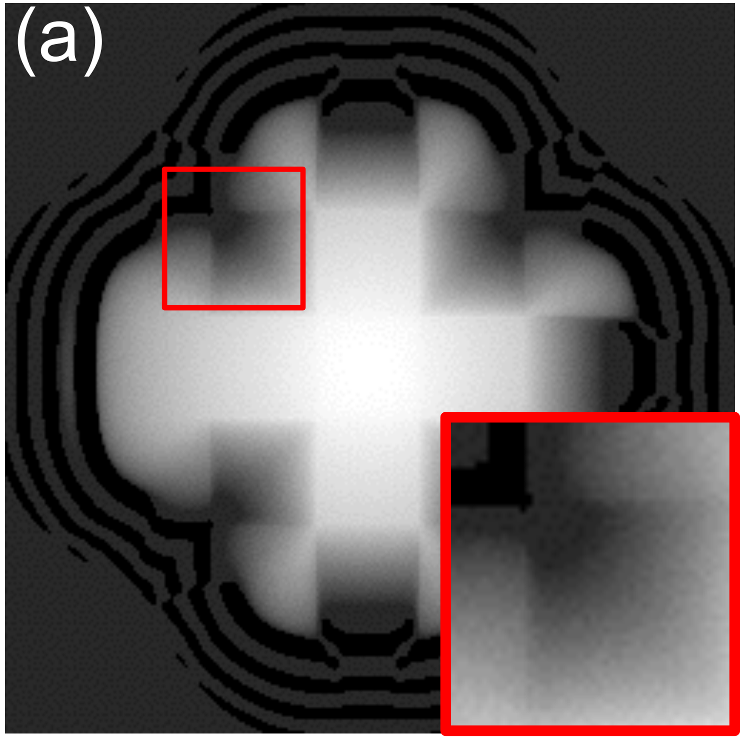


Compressed



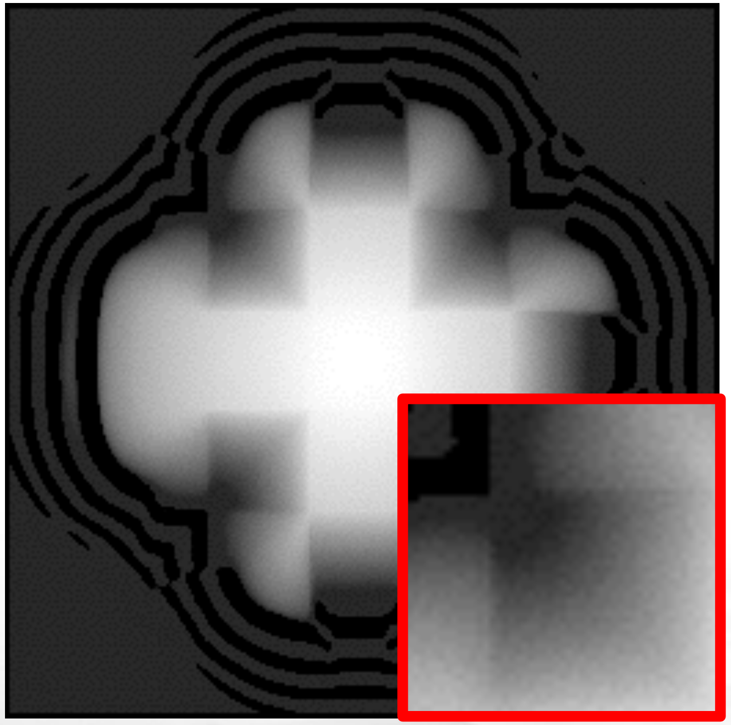
Restored RDN with IBTL

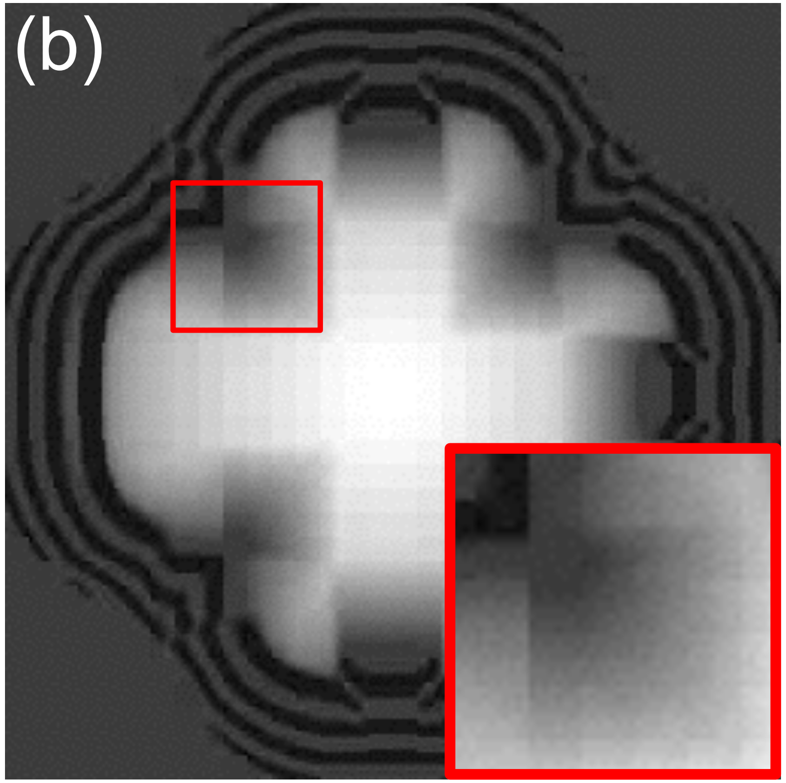
Case Study - Kinetic Fusion Data



Original

Restored RDN with IBTL





Compressed

Less is Really More!

Adaptive Evaluation and Steering of Complex Workflows

BROOKHAVEN
NATIONAL LABORATORY

U.S. DEPARTMENT OF
ENERGY

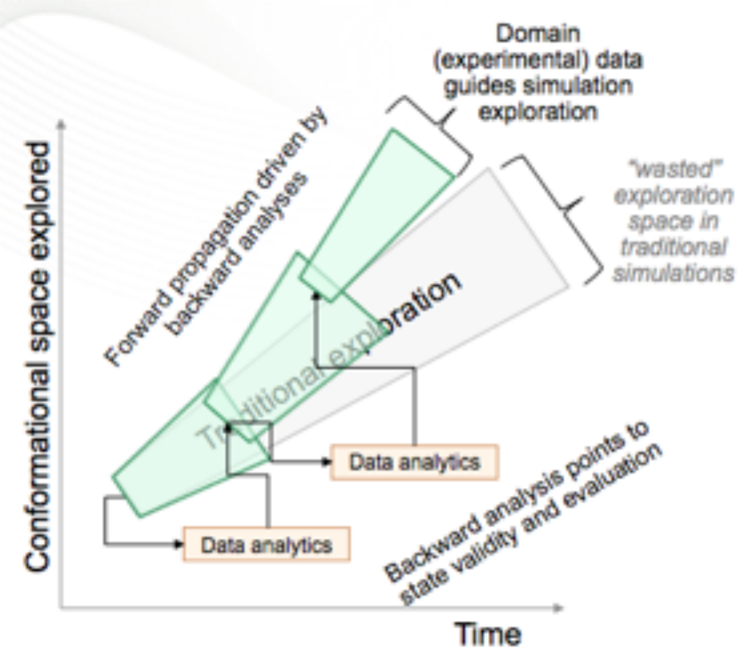
BROOKHAVEN SCIENCE ASSOCIATES

Adaptive Ensemble Algorithms

- Generate ensemble of simulations in parallel as opposed to one realization of process
  - Statistical approach: **O(10^6 - 10^8)** !
  - e.g. Chemistry, Biology, Climate

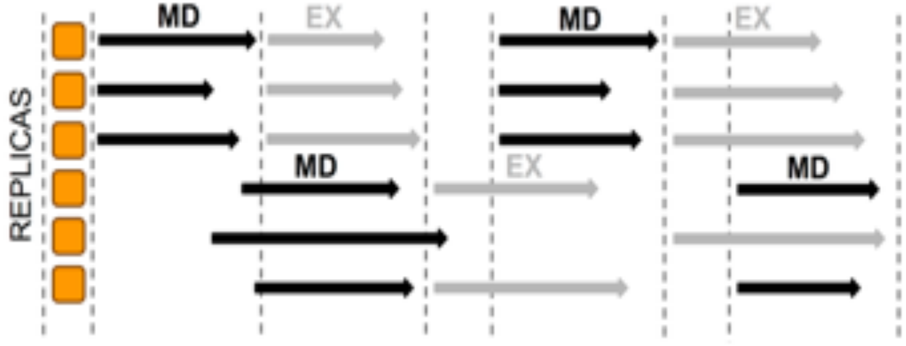
- Ensemble methods necessary, not sufficient!
  - Adaptive Ensembles: Intermediate data, determines next stages

- **Adaptive Ensemble simulations** can easily be **2-3 orders of magnitude faster** than vanilla

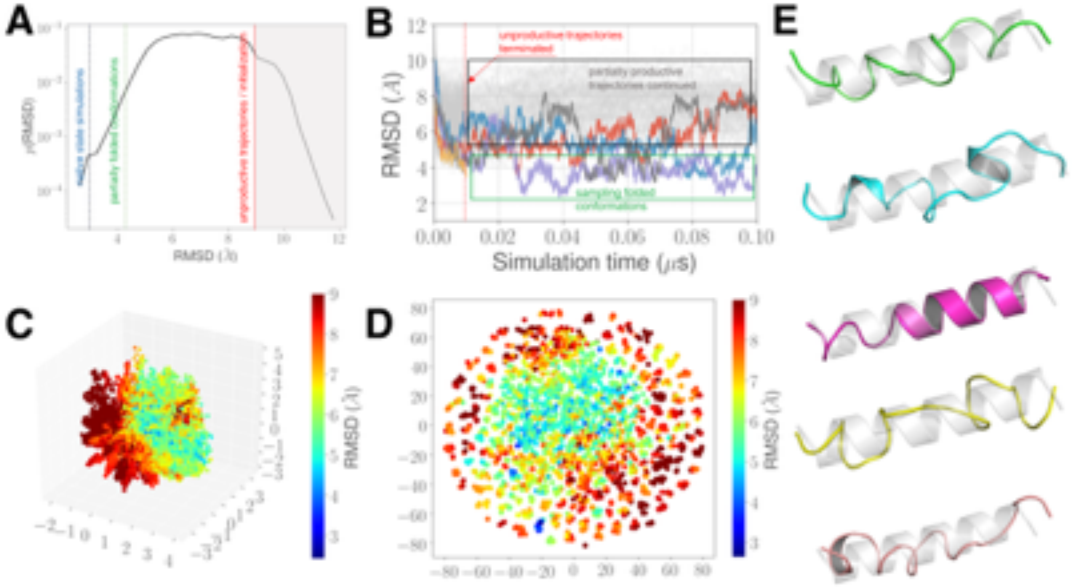


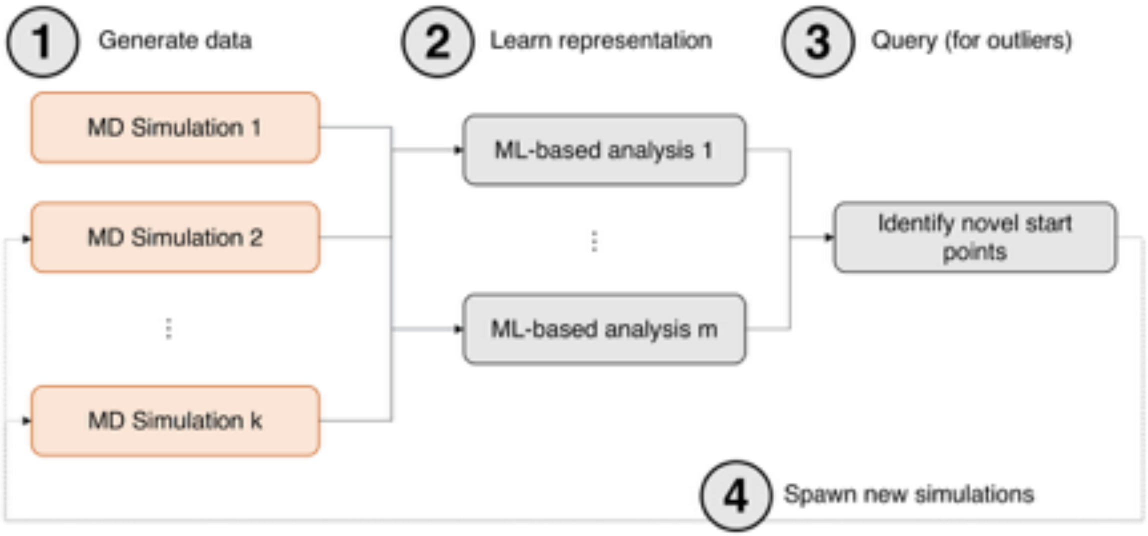
Domain (experimental) data guides simulation exploration
"wasted" exploration space in traditional simulations
Forward propagation driven by backward analyses
Traditional exploration
Data analytics
Data analytics
Backward analysis points to state validity and evaluation
Conformational space explored
Time

*Chodera, J.D., Noe, F., Curr. Opin. Struct. Biol. (2014)



DeepDriveMD https://arxiv.org/abs/1909.07817





1 Generate data

2 Learn representation

3 Query (for outliers)

MD Simulation 1

MD Simulation 2

⋮

MD Simulation k

ML-based analysis 1

⋮

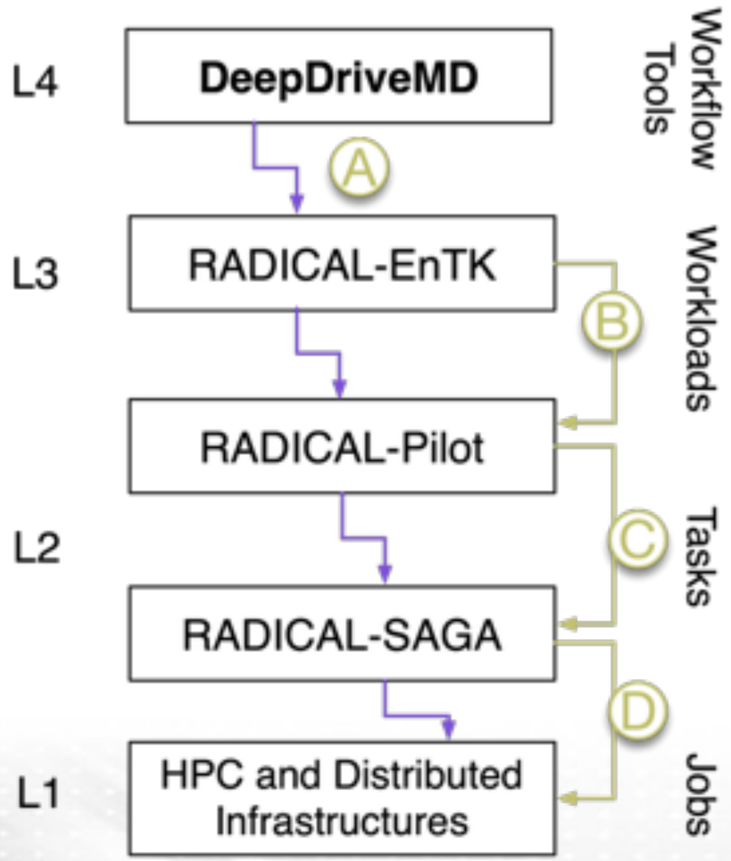
ML-based analysis m

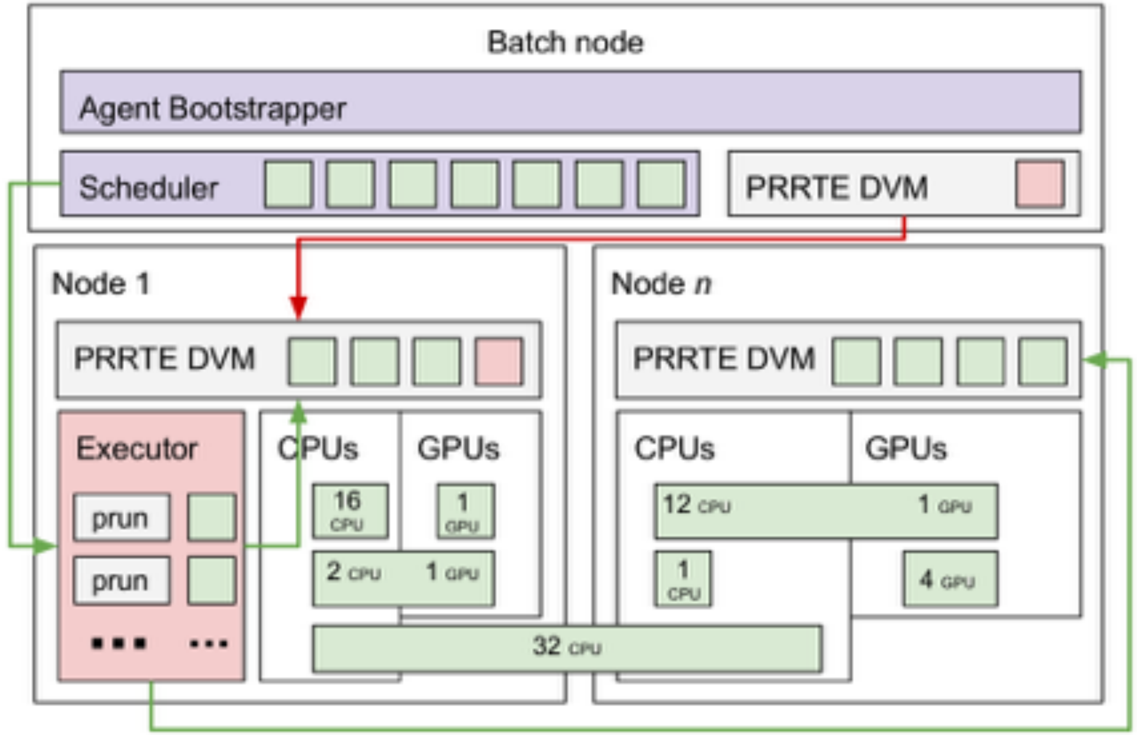
Identify novel start points

4 Spawn new simulations

- Protein Folding using ML driven MD
  - ML Model: Convolutional Variational Auto Encoders (CVAE)
  - CVAE ingests **intermediate data, determines next stages**

RADICAL: Managing ML & HPC tasks on Summit





RADICAL & DeepDriveMD - Results

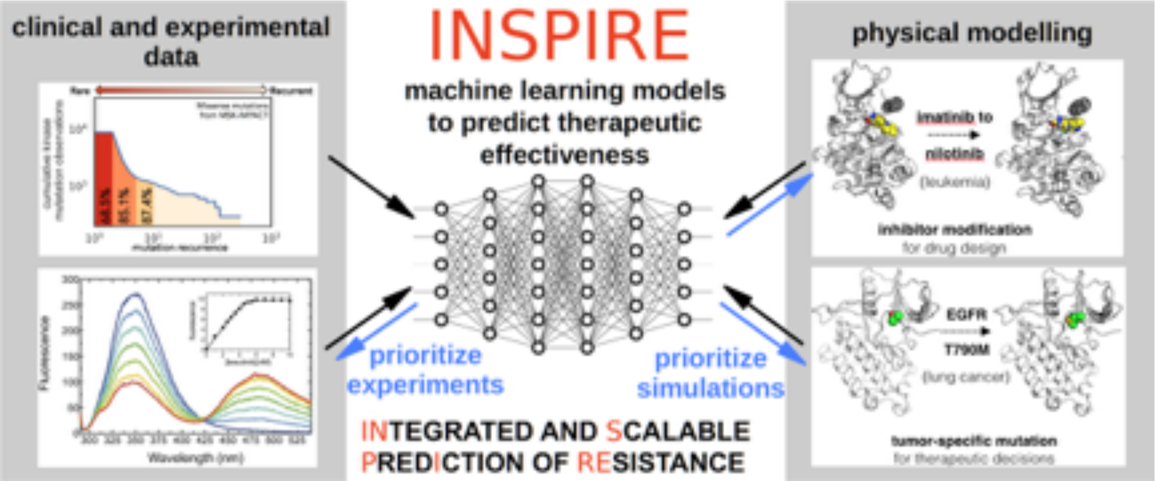
- Manage execution of ensembles
  - Currently: O(**10^3**) on Summit,
  - @ Exascale: **O(10^6 - 10^8)** !
- Concurrent and adaptive training, simulation and inference tasks (right)
- **DeepDriveMD: DL models to adaptively drive ensemble simulations**
- Depending upon data volumes involved:
  - Stream simulation data directly to ML
  - Use “in memory” databases
- **Using DL shows 20x improvement over non-DL adaptive approaches**

RADICAL & Adaptive Ensembles & CANDLE

- Chemical space of drug design in response to mutations very large. 10K -100K mutations; too large for HPC simulations alone!
- Use ML to enhance the **effective performance** of HPC simulations
- Develop methods that use:

  (i) Simulations to train ML models to predict therapeutic effectiveness

  (ii) Use ML models to determine which drug candidates to simulate



clinical and experimental data

INSPIRE

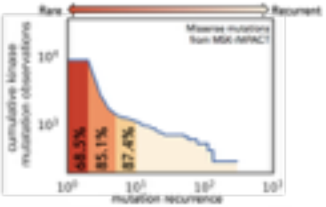
machine learning models to predict therapeutic effectiveness

physical modelling

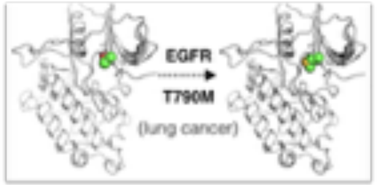
prioritize experiments

prioritize simulations

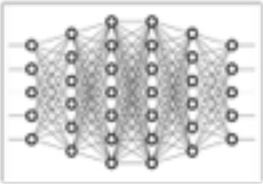
INTEGRATED AND SCALABLE PREDICTION OF RESISTANCE



55,000+ clinical cancer kinase mutations



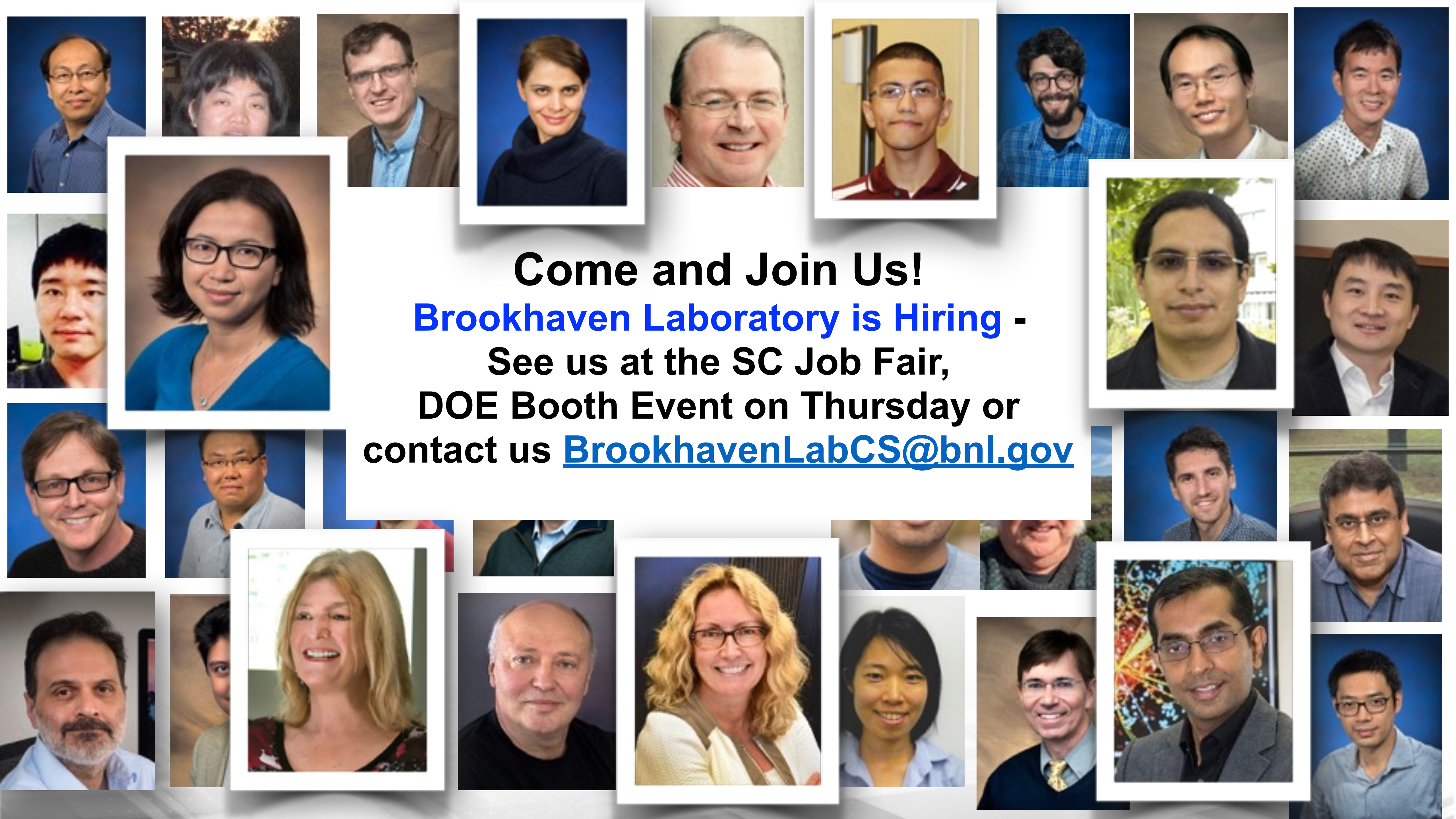
MD simulations to compute kinase inhibitor resistance



Train machine learning (ML) models to predict therapeutic effectiveness

BNL
Brookhaven National Laboratory
OPERATED BY BROOKHAVEN SCIENCE ASSOCIATES UNDER CONTRACT WITH THE UNITED STATES DEPARTMENT OF ENERGY

Thank You!



Come and Join Us!

Brookhaven Laboratory is Hiring -

See us at the SC Job Fair,

DOE Booth Event on Thursday or

contact us BrookhavenLabCS@bnl.gov