ExaLearn Overview
ECP’s Co-design Center for Exascale Machine Learning Technologies

Project PI: Francis J. Alexander, BNL
Partner PIs and Institutions:

- Ian Foster, ANL
- Peter Nugent, LBNL
- Brian Van Essen, LLNL
- Aric Hagberg, LANL
- David Womble, ORNL
- James A. Ang, PNNL
- Michael Wolf, SNL

Place: HPC User Forum, Santa Fe, NM
Date: April 2, 2019

Co-design is a Multidisciplinary Collaborative Process

- In Co-design, HPC systems are developed with application software influencing hardware design trade-offs while also recognizing that applications and supporting software must be developed concurrently with hardware changes.
- Under the status quo, HPC systems are developed without application input, and “advanced architectures” have a reputation of being “hard to program.”
- Co-design encourages all parties, ranging from the system designers, computer architects, application software and tools developers, facilities, etc., to jointly design and optimize the system specifications via open communication and collaboration.

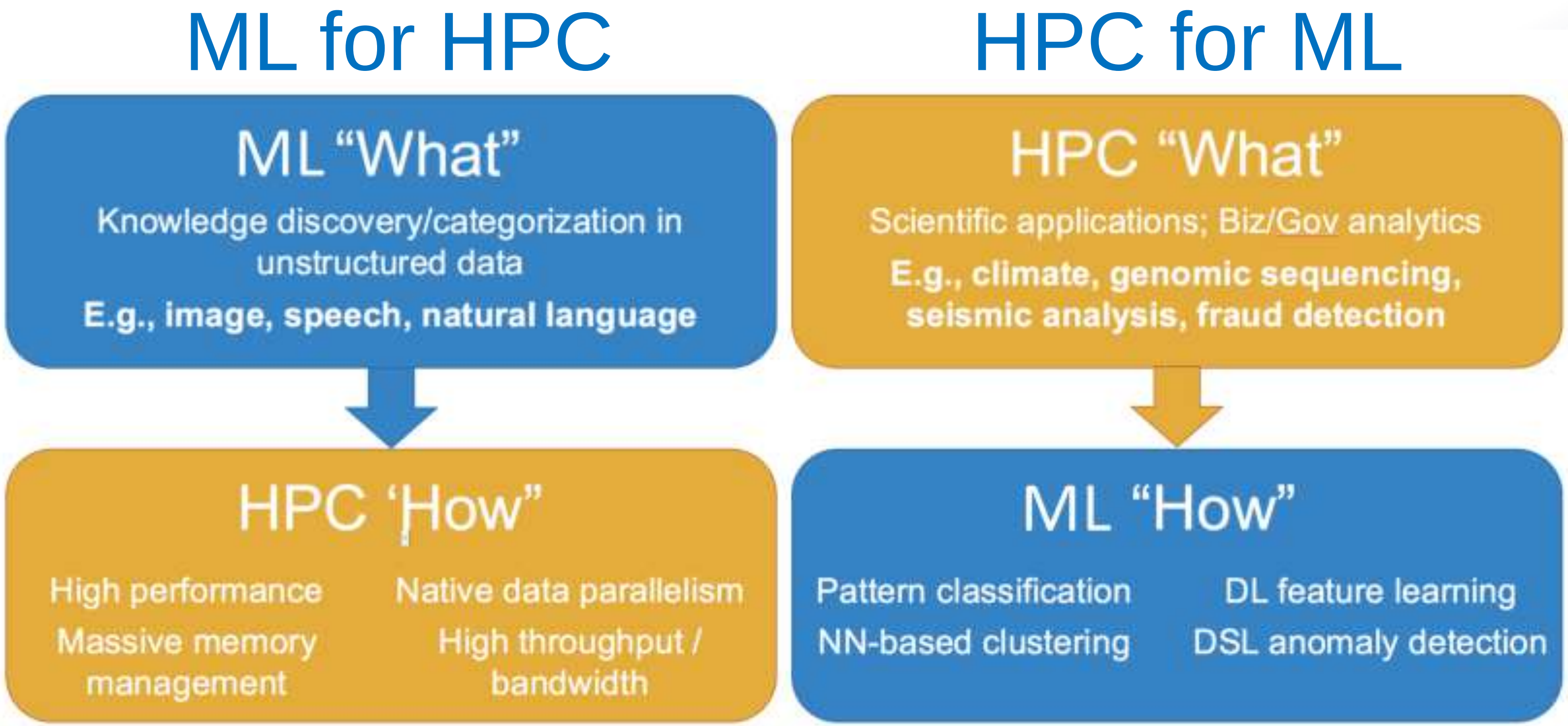
ECP’s Portfolio of Co-design Projects

- CODAR: Co-design center for Online Data Analysis and Reduction
- COPA: Co-design Center for Particle Applications
- AMREX: Block-structured adaptive mesh refinement (AMR) Co-design Center
- CEED: Center for Efficient Exascale Discretizations (CEED)
- ExaGraph: Combinatorial Methods for Enabling Exascale Applications
- ExaLearn: Co-design Center for Exascale Machine Learning Technologies

Three Overarching ExaLearn Co-design Questions

- What are the best learning algorithms and methods for different application classes in terms of speed, accuracy, and resource requirements? How can we implement these algorithms to achieve scalability and performance portability?
- What are the trade-offs in learning accuracy, resource needs, and overall application performance between using various learning methods? How do these trade-offs vary with exascale hardware and software choices?
- How do we effectively orchestrate learning to reduce associated overheads? How can exascale hardware and software help with this orchestration?

Relationships between Machine Learning and HPC



Ack: Gadi Singer

- HPC for Machine Learning: HPC Technologies are applied to learning tasks to accelerate computation and/or solve larger problems.
- Machine Learning for HPC: Learning Technologies are applied to HPC computations to improve their performance in some way, e.g., by choosing the next simulation(s) to perform.

Application Priorities Determine Machine Learning Methods

- Deep Learning (CNN, RNN, etc.)
- Ensemble Methods and Random Forest Methods
- Reinforcement Learning
- Kernel Methods
- Tensor Methods
- Graph-Based learning

Overarching Goals for ExaLearn

- Provide exascale machine learning software for use by:
  - ECP Applications Projects
  - Other ECP Co-design Centers
  - DOE Experimental Facilities
  - DOE Leadership Class Computing Facilities
- ExaLearn will establish multidisciplinary collaborations in learning technologies that cross-cut individual ECP projects:
  - ECP AD projects that share an interest in a machine learning method
  - ECP ST projects
  - ECP HI/PathForward projects

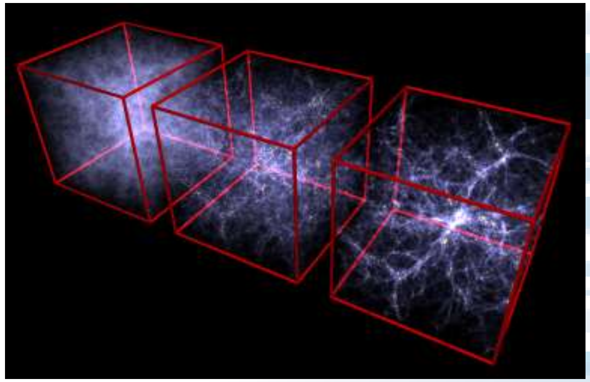
Work Product is a Software Toolset that…

- Applies to multiple problems within the DOE mission
- Has a line-of-sight to exascale computing, e.g., by using exascale platforms directly or providing essential components to an exascale workflow
- Does not replicate capabilities easily obtainable from existing, widely available packages
- Builds in domain knowledge where possible (not often done by industry, although efforts beginning IBM, GE, etc.); “physics”-based ML and AI (recurring theme)
- Quantifies Uncertainty in predictive capacity
- Is interpretable
- Is reproducible.

Application Evaluation Criteria

- Would the work we propose directly benefit an ECP application/program, and/or an important non-ECP application, and/or DOE user facility? Impact could include facility design and control, design of experiment, etc.
- Is it exascale in size?
  (a) i.e., tuning hyperparameters on individual nodes over the whole machine (pseudo-embarrassingly parallel)
  (b) i.e., deploying a complex network across the whole machine
  (c) does any aspect of the ML problem (either data size or optimization time, etc., require exascale)
- Does it enable new science? What is the specific science question?
- What ML methods do they employ right now?
- Do they have examples in-hand on which we can benchmark at scale now?
- Are there cross-cut benefits to other science areas?
- What are the known performance considerations and/or tool barriers ?
- What data are available, what is the quality of the data, how is it curated, and what are the risks associated with using that data?
- What would the workflow look like (training/ensemble/optimization/etc.)?
- Is there willingness/interest on the part of the application/facilities folks in working with ExaLearn ?
- Is there any existing software code base/how will feedback work/can we make a meaningful impact/will it lead to future work like SciDAC?

ExaLearn for Surrogates–Making Realistic Simulations on the *Cheap*



- **Challenge and Importance**: Many DOE simulation efforts could benefit from having realistic surrogate models in place of computationally expensive simulations. These can be used to quickly flesh out parameter space, help with real-time decision making, assist experimental design, and determine the best areas to perform additional simulations. We will first target large-scale structure simulations of the universe. As the field is well developed, the scale can easily be ramped up to an exascale ML challenge, and the field is robust enough to explore systematics at the sub-percent level.
- **ML impact**: Neural-networks-based generative models can make reliable surrogate models of expensive simulations for data augmentation purposes. Such surrogate models can be used to aid in cosmological analysis to reduce systematic uncertainties in observations.
- **Timeliness**: The ExaSky Application projects is producing the largest LSS simulations now. The DESI experiment starts next year, and LSST takes its first science images in 2021.
- **Urgency**: All cosmological measurements today are limited by systematics, not statistics. To reduce these uncertainties and make the most of these future experiments, thousands, if not millions, of exascale-sized simulations must be carried out. Surrogate models are a viable path forward to achieving this goal—only if their limitations are fully understood.
- **Benefit to ECP–Large DOE Experiments**: Once demonstrated, this software framework can be easily adapted to other fields and simulation areas.

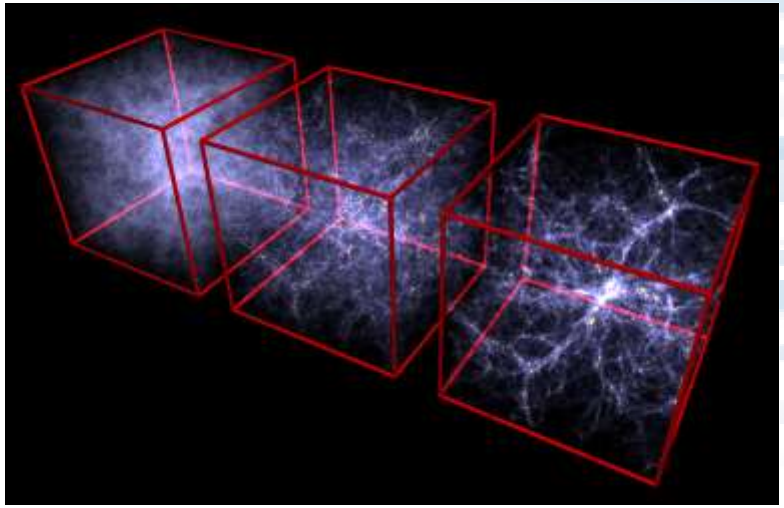
EXASCALE COMPUTING PROJECT

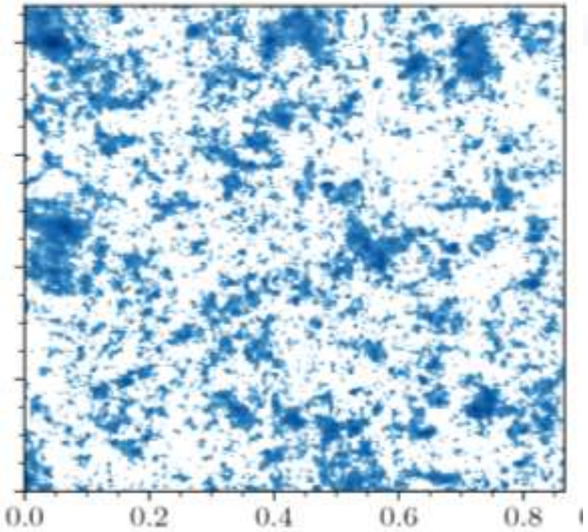
CosmoGAN for Surrogates

Mustafa Mustafa, Deborah Bard, Wahid Bhimji,
Rami Al-Rfou, Zarija Lukić
https://arxiv.org/abs/1706.02390

- Cosmology simulations are extremely computationally expensive:
  - e.g., ‘Q-continuum’ HACC sim took ~2 weeks
  - One product is weak lensing convergence maps to compare to observations
- Seek to augment simulations with a learned generative neural net
  - Reproduce convergence maps–images
  - Train using existing simulation maps with different cosmological parameters
  - Once trained produce new examples
  - Use these maps to constrain observational systematics quickly!

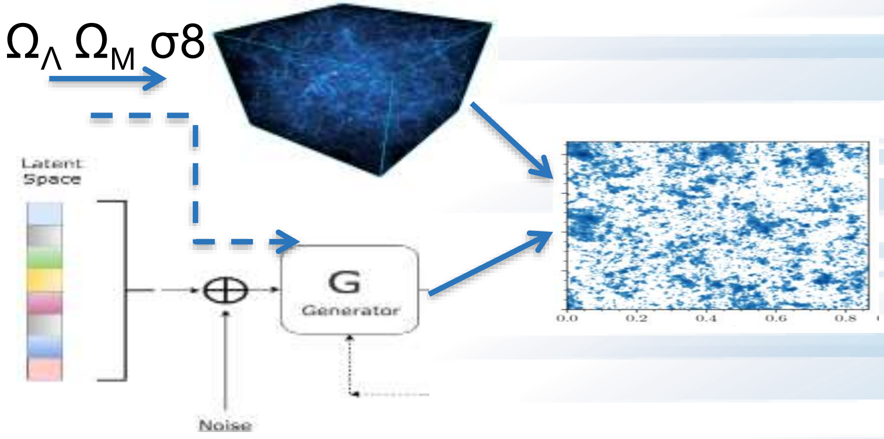
11





EXASCALE COMPUTING PROJECT

GANs and Cosmology

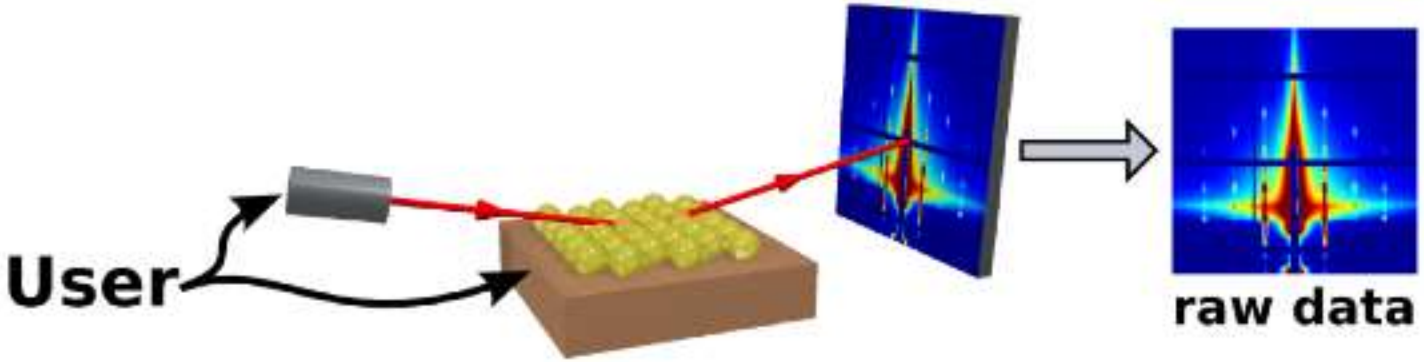


Exascale in size:

- The surrogate model used are based on the Generative Adversarial Networks (GAN) framework and currently are part of experiments with Variational Auto-encoders. Both models will eventually need hyperparameter tuning across an exascale-sized machine to extend to larger maps and build conditional generative models.
- For complex conditional models on large maps (2048^2 with the expectation to go to 8096^3), we expect to use intra-node model parallelism (up to a few GPUs) and inter-node data parallelism (10s-100s of nodes) for faster experimentation during algorithm development and final model training. When such models' hyperparameters are being tuned, they will require thousands of nodes.
- After having been trained, when applied to generation, this GAN/VAE approach would be coupled with large-scale cosmology applications running on exascale machines (such as the ExaSky Apps HACC and Nyx).
- Future observational data sets will be multi-petabyte in size, and the simulation data are larger by an order of magnitude or more.

Machine Learning for Control: Directed Self-Assembly of Block Copolymers

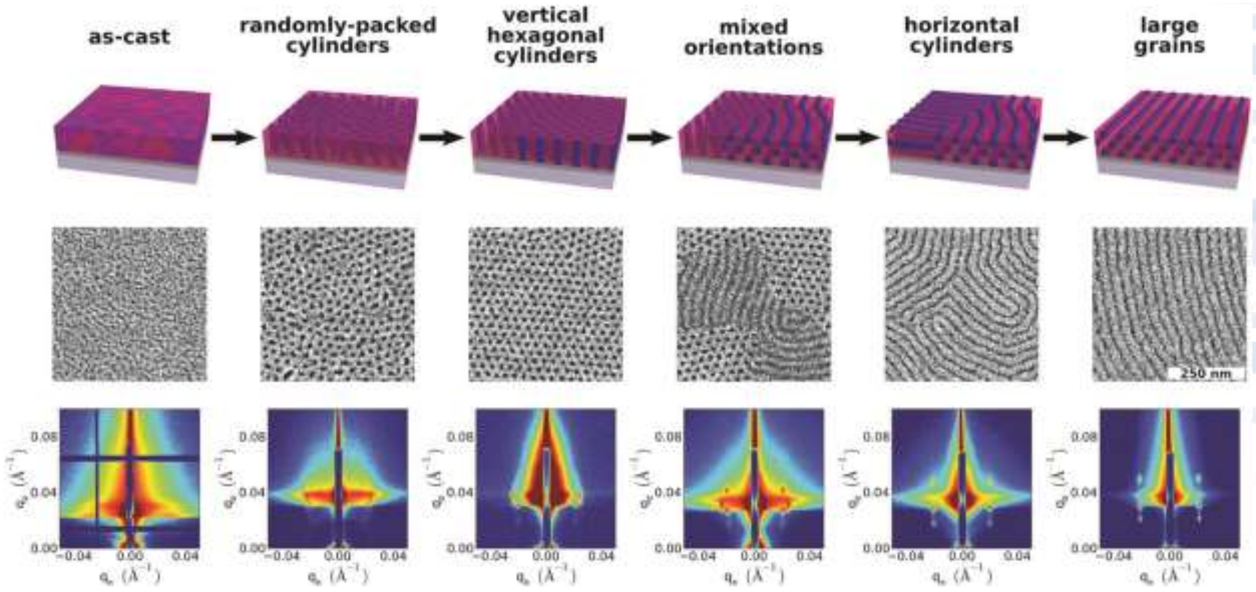
- **Science driver:** Thin films of block-copolymer (BCP) cylinder-forming polystyrene-block-poly(methyl methacrylate) are self-assembled during experiments at light sources. This is important to the understanding and annealing of materials at the nanoscale.
- **Our Use Case:** Use reinforcement learning to develop a policy to guide the scientist to modify experimental parameters (e.g., temperature) most likely to affect outcomes favorably with the goal of forming non-equilibrium morphologies.
- **Experiment:** Grazing-Incidence Small-Angle X-ray Scattering (GISAXS) experiments can determine symmetry, size, spacing, orientation, grain size, order/perfection, unit cell, and more.



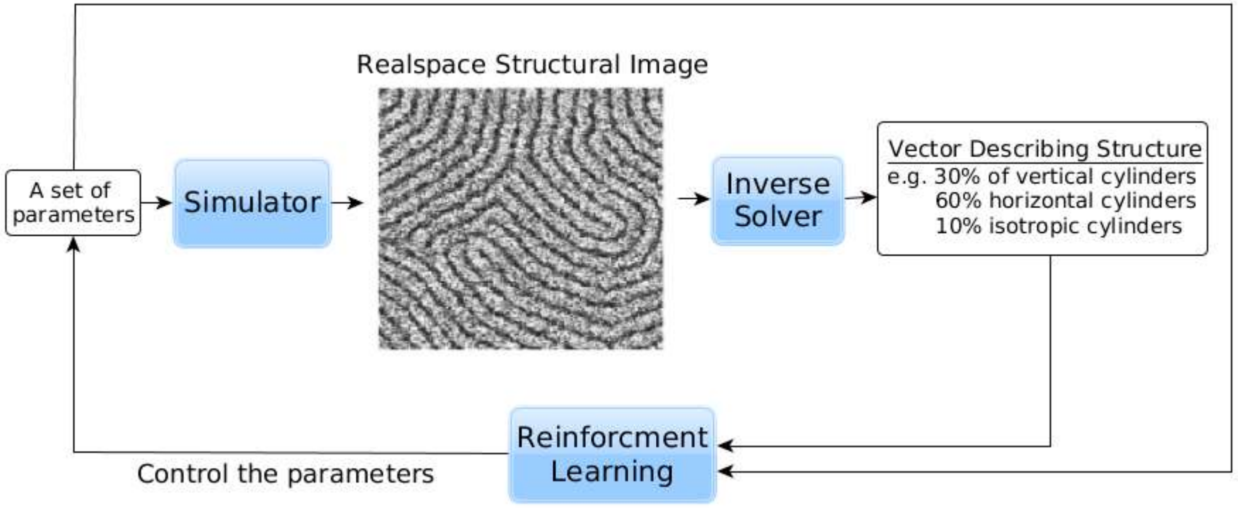
- **Synchrotron light sources** are used for GISAXS experiments.
- **Challenge:** Accelerate self-assembly process with minimal use of experimental and computational resources.
- **Approach:** Train reinforcement learning model at supercomputing facility and deploy at user facility. Use synthetic data for initial training and eventually incorporate experiment data into the model.

BCP Problem and Initial Reinforcement Learning Workflow Design

- BCP cylinders go from poorly ordered to hexagonal then to horizontally ordered.
- The size, shape, and anisotropy of grains are influenced by ordering history; for instance, faster heating rates reduce grain anisotropy (different properties in different directions).



Cartoon, SEM and GISAXS images of BCP ordering history (from P. Majewski and K. Yager)

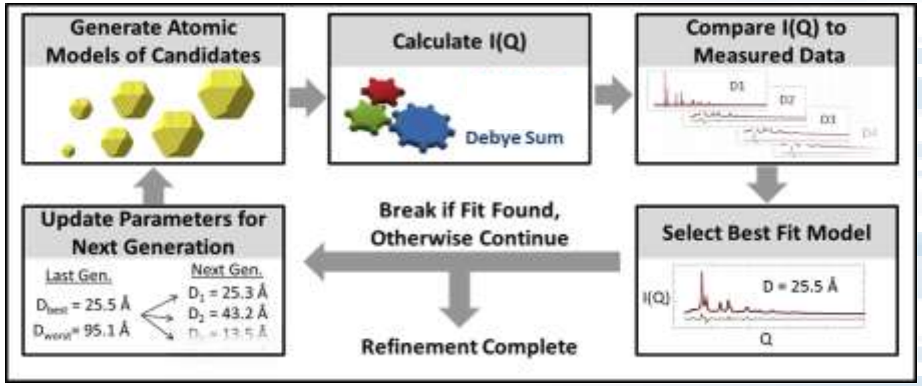


- Reinforcement learning system will first train on simulated images that are converted to a structure vector.
- Control parameters will be annealing temperature and time, BCO composition, thin film thickness, substrate surface chemistry, blending ratio of different BCPs, and more.

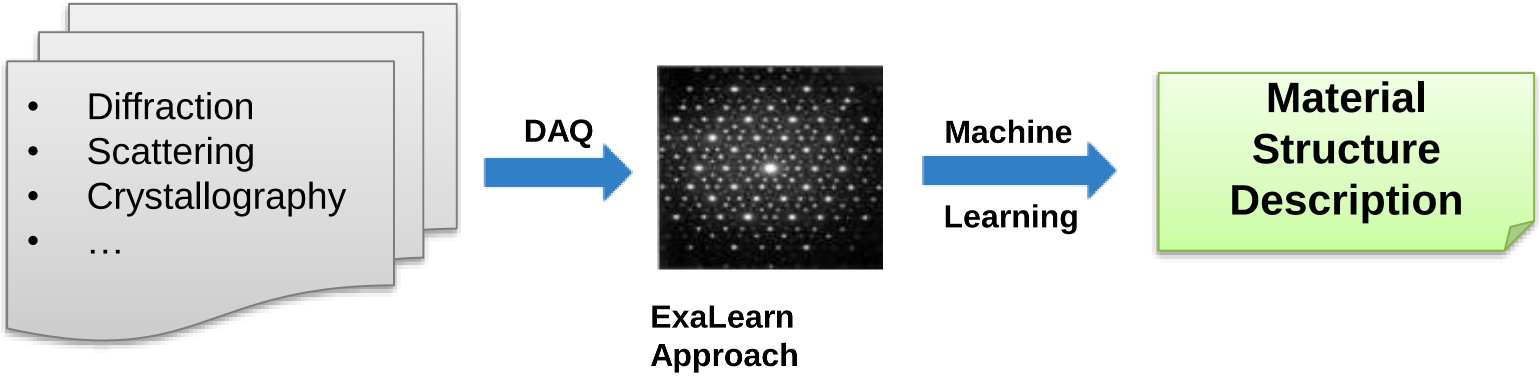
EXASCALE COMPUTING PROJECT

ExaLearn: Machine Learning for Inverse Problems in Materials Science

- **Objective**
  - Experimental data acquisition (DAQ) technologies
    - Diffraction
    - Scattering
    - Crystallography
    - Other experimental procedures
  - Current methods use forward models in refinement loops for determination of material structure and properties
    - Extremely time-consuming
    - Speed, correctness, and quality of results depend on physics included in forward models
  - ExaLearn will architect and deploy extreme-scale ML tools to enable fast, accurate determination of materials structure ***directly*** from experimental data.

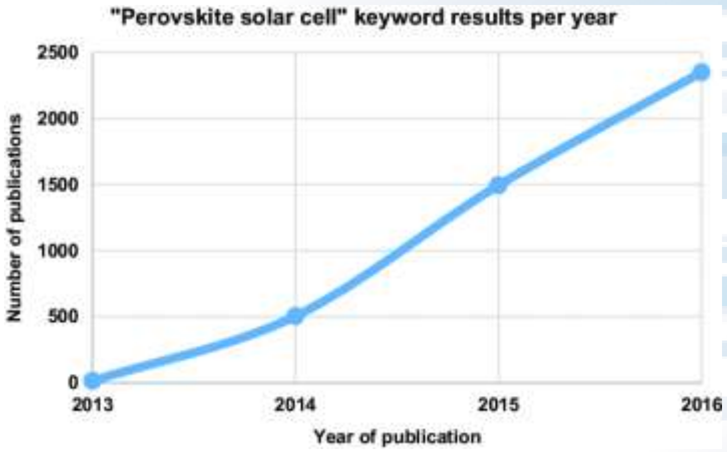


Expensive Loop Refinement Method

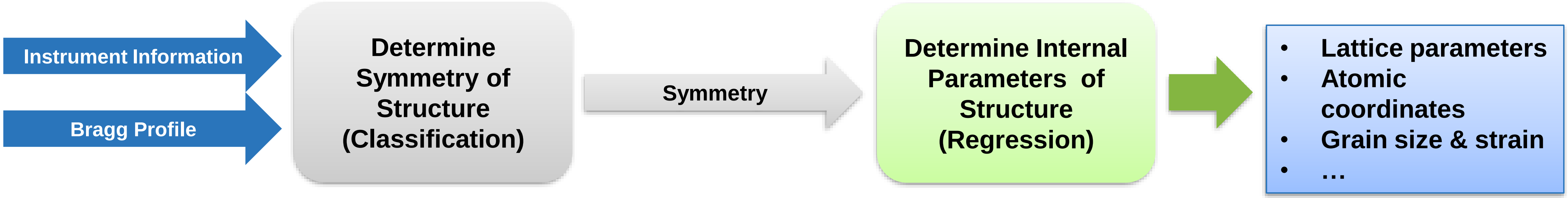


ExaLearn: Machine Learning for Inverse Problems in Materials Science

- **Impact**
  - Enable learning from vast troves of materials-related empirical and simulation data available within the DOE complex, as well as in various repositories around the world ***quickly*** and ***more accurately***.
  - Core ML technologies will be built based on ***domain-agnostic design principles***, but deployment will target domain-specific benchmarks in *neutron scattering* and *X-ray crystallography*.
  - Of particular interest are studies of an important class of materials called ***perovskites*** (see chart).



ExaLearn Pipeline for Material Structure Determination from Neutron Scattering/Diffraction Data

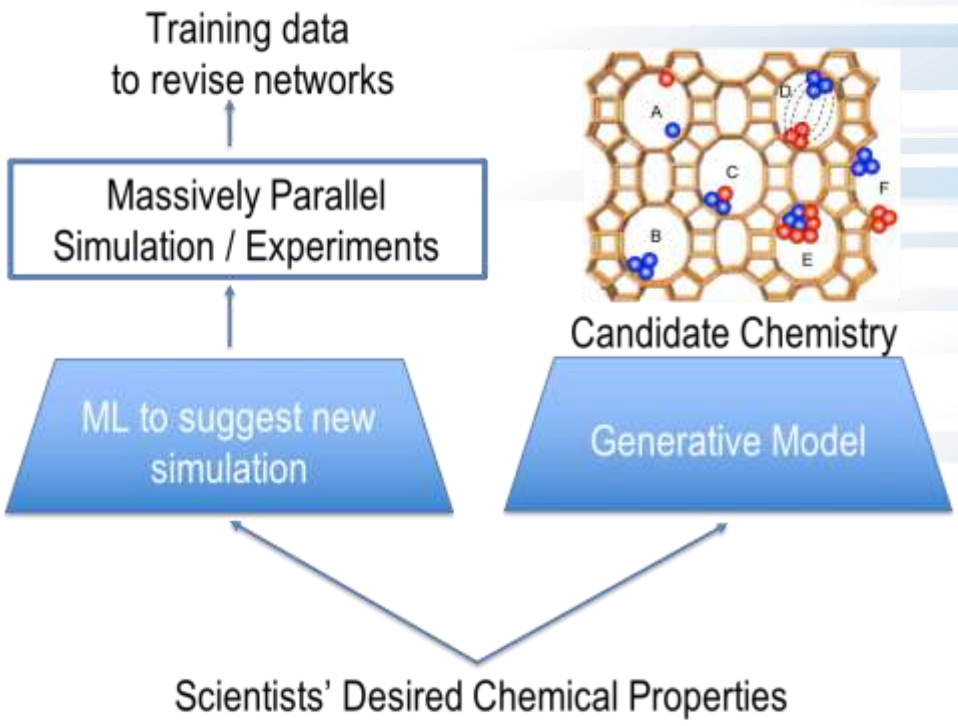


- **Point of Contact** Sudip K. Seal (sealsk@ornl.gov, 865-574-8152)

Perovskite solar cells: An integrated hybrid lifecycle assessment and review in comparison with other photovoltaic technologies - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Number-of-publications-resulting-from-the-search-of-perovskite-solar-cell-on-Web-of_fig3_317569861 [accessed 11 Jan, 2019]

ExaLearn: Machine Learning for Design–Goal

- Find novel materials with beneficial properties.
- Combine simulation and reinforcement learning at exascale to learn to intelligently explore design space.
- Use simulation data to train generative models to propose new materials for scientists to test.
- **Use Case 1: Electrolytes for improved batteries and energy storage.**
- **Use Case 2: Catalysis for improved chemical conversions and transformations.**



Scaling and Performance Cross-cut Area

- The scaling and performance cross-cut area's goal is to identify or create supporting ML/deep learning (DL) technologies and tools that integrate into the software and infrastructure cross-cut area and serve the research needs for the four application areas.
- The four application areas have diverse ML/DL requirements ranging from:
  - Classification, regression, and generative DL models
  - Reinforcement learning
  - ML/DL-guided workflows driving predictive science applications
    - Lightweight, *in situ* inference
    - Continuous integration of new simulation output into trained model
  - Creating fast surrogate models

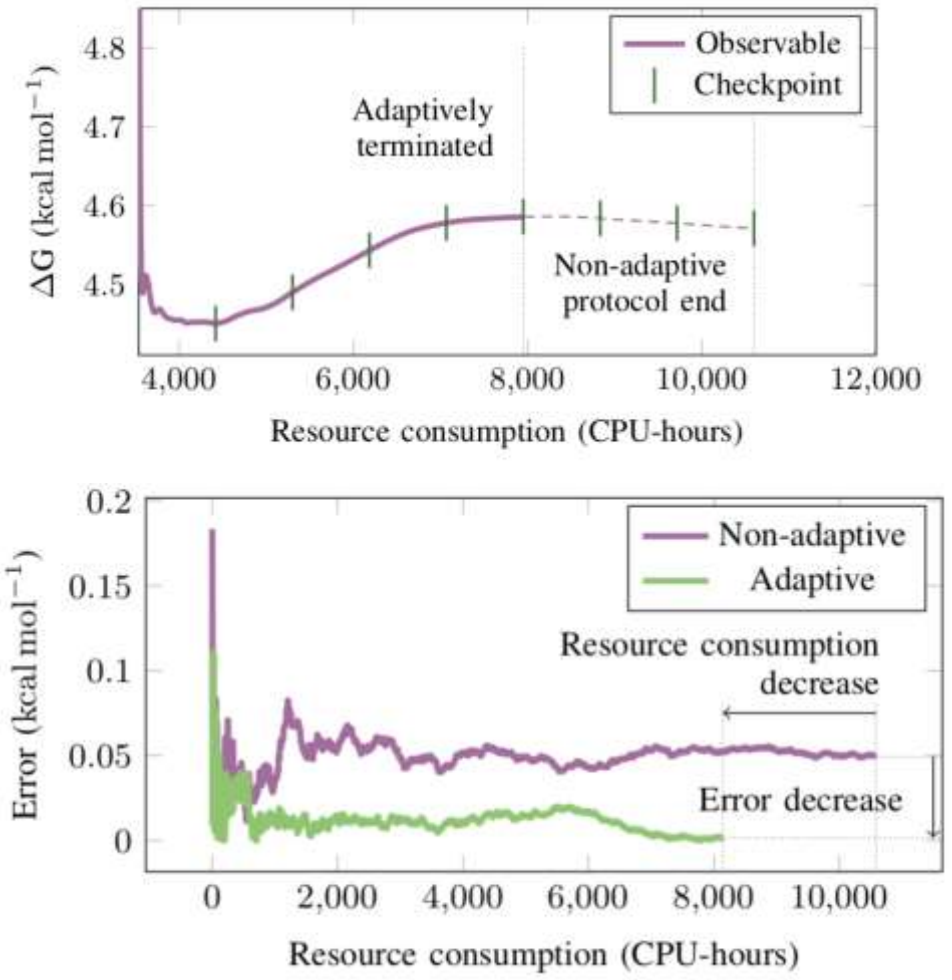
Adapting Scalable ML/DL tools to ExaLearn Applications

- Application needs will be served by a combination of existing open-source research tools, for example:
  - Horovod plus TensorFlow and PyTorch
  - LBANN
  - OpenAI Gym
  - CANDLE supervisor framework

Tool	Data Parallelism	Model Parallelism	Ensemble Parallelism	Environment parallelism
TensorFlow + Horovod	X			
PyTorch + Horovod	X			
LBANN	Y	Y	Y (LTFB)	
OpenAI Gym				Y
CANDLE Sup.			Y (PPBT)	Y

Objective-driven Experiment Design (ODED)

- ODED: Intelligently steer a computational campaign toward an objective. Determine “optimal” next computing steps dynamically. A fundamental dependence on adaptive workflows.
- ODED is a common functional requirement across all four ExaLearn applications: Control, Design, Inverse, and Surrogates. Also a recurring requirement in Climate Modeling, Computational Drug Design (CANDLE)
- Exemplar: ODED for Drug Design. Given a set of drug candidates and computationally expensive protocols, find optimal mix of protocols and parameters given a constraint in HPC resource/allocation/time-to-solution.
- **ExaLearn ODED Framework** (Brookhaven Lead)



dlhub.org

DLHub

Data and Learning Hub for Science

A simple way to find, share, publish, and run machine learning models and discover training data for science

Documentation

Read the Docs | Python SDK | CLI

Papers and Presentations

DLHub on ArXiV | DLHub Slides

Get Started

1

Describe

```
m = KerasModel()
m.create_model("p1b1-example.h5")

m.set_title("CANDLE Pilot 1 - Benchmark 1")
m.set_name("candle_p1b1") # short name
m.set_domains("genomics","biology","HPC")
```

2

Publish

```
from dlhub_sdk.client import DLHubClient

dl = DLHubClient()
dl.publish_servable(m)
```

3

Run

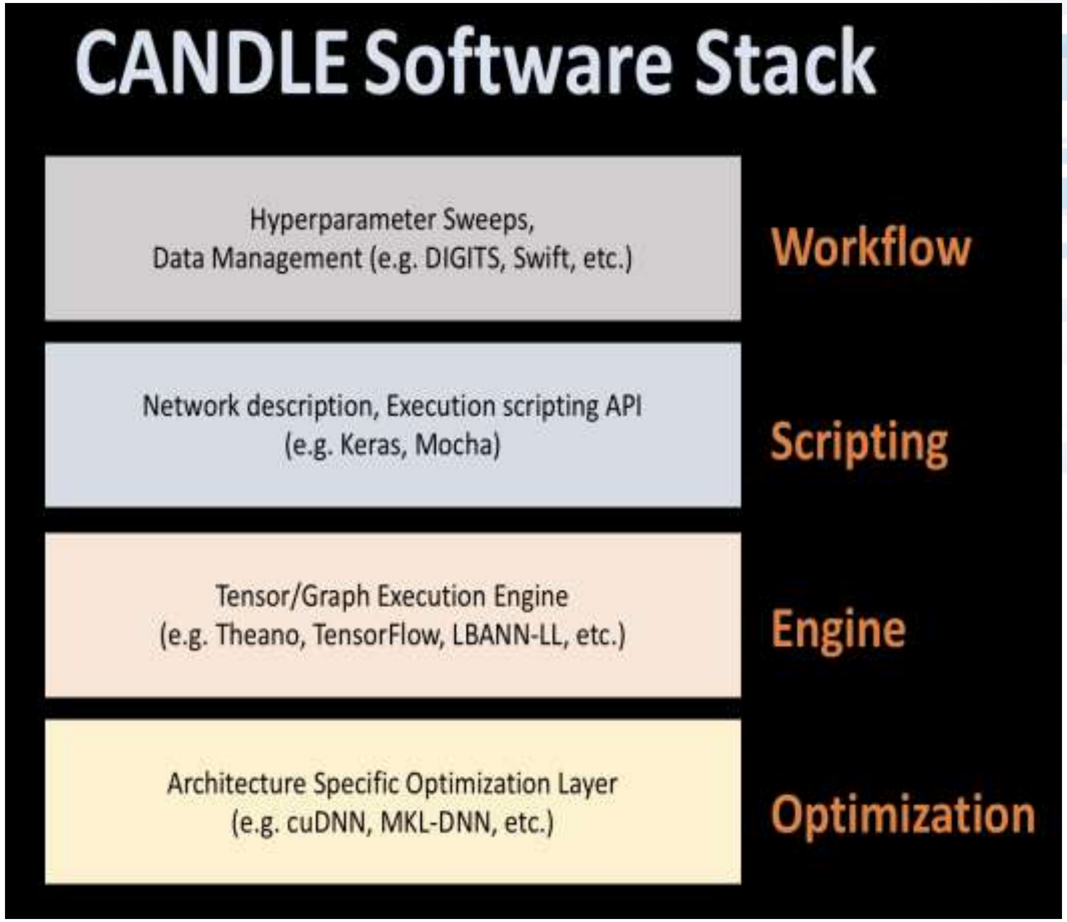
```
from dlhub_sdk.client import DLHubClient

dl = DLHubClient()

mid = dl.get_id_by_name("candle_p1b1")
data = np.load("pilot1.npy")
pred = dl.run(mid, data.tolist())
```

ExaLearn ECP PathForward Vendor Engagement

- Leads: Michael Wolf (SNL) and James Ang (PNNL)
- Three major efforts
  - **Direct vendor engagement** (Cray, Intel, HPE, AMD, IBM, NVIDIA). Discuss ExaLearn needs with vendors, attend vendor “deep-dives,” and understand new technologies.
  - **ECP application and experimental facility engagement**. Work with other seven ExaLearn application and cross-cutting areas to develop proxy applications to reflect the system utilization (computation, communication, etc.) of key ML kernels accurately in the context of applications.
  - **ECP ML proxy applications**. Work with ECP proxy application team to define ExaLearn proxy applications and problem sets as part of ECP proxy application suite. Work with ECP application development teams and PathForward vendors to ensure these are useful ML drivers.

CANDLE Project Provides Initial Take on Exascale Learning Software Requirements

- Enable high productivity for DL-centric workflows
- Support Key DL frameworks on DOE supercomputers (Keras, TF, Mxnet, CNTK)
- Support multiple paths to concurrency (Ensembles, Data and Model Parallel)
- Manage training data, model search, scoring, optimization, production training and inference (End-to-End Workflow)
- CANDLE runtime/supervisor (interface with batch schedulers)
- CANDLE Python library for improving model development (UQ, HPO, CV, MV)
- Well-documented open examples and tutorials on GitHub
- Leverage as much open source as possible (build only what we need to add to existing frameworks)



ExaLearn will fill gaps, address new requirements, support other forms of learning, address usability and scalability, and otherwise meet needs of new applications

Thank you!