AI for Science

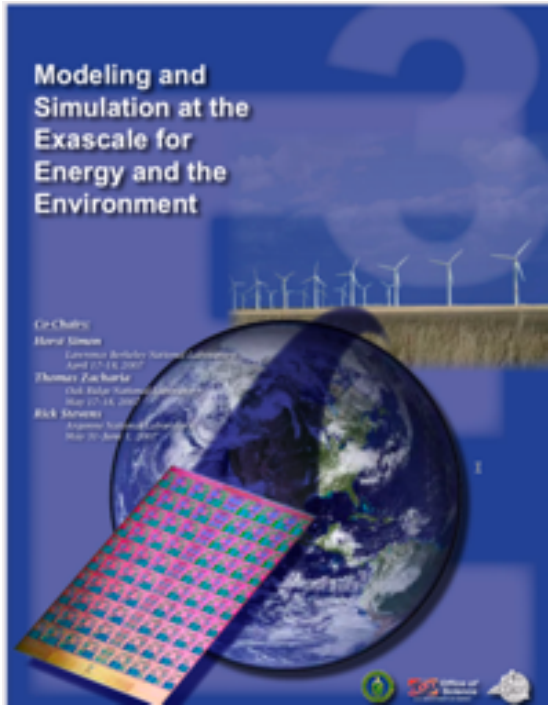
Rick Stevens
Argonne National Laboratory
The University of Chicago

Crescat scientia; vita excolatur

AI for Science Townhalls

Organized by Argonne, Oak Ridge and Berkeley with participation from all the laboratories..

- Four “Townhalls” aimed at getting input from the DOE community on opportunities and requirements for the next 5-10 years in computing with a focus on convergence between HPC and AI
- July (Argonne), August (Oak Ridge), September (Berkeley), October (Washington)
- Modeled after the 2007 Townhalls that launched the Exascale Computing Initiative
- Each meeting covers roughly the same ground, geographically distributed to enable local participation
- Applications in science, energy and technology
- Software, math and methods, hardware, data management, computing facilities, infrastructure, integration with experimental facilities, etc.
- Expect ~200 people per meeting
- Output will be a report to guide strategic planning at Labs and DOE



What We are on About at the Townhalls

- AI is transforming our “regular life” world
- AI has tremendous potential to accelerate scientific discovery
- How do we go about organizing an AI for Science initiative
- Capture ideas, problems, requirements and challenges for an AI for Science initiative
- What problems could be attacked?
- What data, simulations, and experiments do we need?
- What kind of methods, software and math do we need?
- What kind of computer architectures and infrastructure do we need?

AI complements our Exascale plans

- The emerging platforms at the LCF and NERSC will be excellent platforms for machine learning, in particular deep learning training
- The coupling of AI and HPC is a huge opportunity for DOE
- Many uses of AI couple to experiments in ways that traditional modeling and simulation do not
- The DOE experimental community could become major users of the DOE HPC facilities
- AI has the potential to accelerate science at all scales
- Future systems directions will be impacted by AI use cases

White House AI Executive Order

EXECUTIVE ORDERS

Executive Order on Maintaining American Leadership in Artificial Intelligence

INFRASTRUCTURE & TECHNOLOGY

Issued on: February 11, 2019

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Policy and Principles. Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life. The United States is the world leader in AI research and development (R&D) and deployment. Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities. The Federal Government plays an important role in facilitating AI R&D, promoting the trust of the American people in the development and deployment of AI-related technologies, training a workforce capable of using AI in their occupations, and protecting the American AI technology base from attempted acquisition by strategic competitors and adversarial nations. Maintaining American leadership in AI requires a concerted effort to promote advancements in technology and innovation, while protecting American technology, economic and national security, civil liberties, privacy, and American values and enhancing international and industry collaboration with foreign partners and allies. It is the policy of the United States Government to sustain and enhance the scientific, technological, and economic leadership position of the United States in AI R&D and deployment through a coordinated Federal Government strategy,

Policy Statement

- Artificial Intelligence (AI) promises to drive growth of the United States economy, enhance our economic and national security, and improve our quality of life.
- Continued American leadership in AI is of paramount importance to maintaining the economic and national security of the United States and to shaping the global evolution of AI in a manner consistent with our Nation's values, policies, and priorities.
- Maintaining American leadership in AI requires a concerted effort to promote advancements in technology and innovation, while protecting American technology, economic and national security, civil liberties, privacy, and American values and **enhancing international and industry collaboration with foreign partners and allies**.

Five Principles

- **Drive technological breakthroughs** in AI across the Federal Government, industry, and academia in order to promote scientific discovery, economic competitiveness, and national security.
- **Drive development of appropriate technical standards** and reduce barriers to the safe testing and deployment of AI technologies in order to enable the creation of new AI-related industries and the adoption of AI by today's industries.
- **Train current and future generations** of American workers with the skills to develop and apply AI technologies to prepare them for today's economy and jobs of the future.
- **Foster public trust and confidence in AI technologies** and protect civil liberties, privacy, and American values in their application in order to fully realize the potential of AI technologies for the American people.
- **Promote an international environment that supports American AI** research and innovation and opens markets for American AI industries, while protecting our technological advantage in AI and protecting our critical AI technologies from acquisition by strategic competitors and adversarial nations.

What is possible ?

Things we can do in Science with AI now

Learn predictive models from data without relying upon theory or deep mechanistic understanding

E*xample: predicting materials and chemistry properties*

Learn approximate solutions to inverse problems where we have data and models are not available or are inefficient

Example: phase retrieval in coherent x-ray imaging

Generate large collections of synthetic data that models real data

Example: synthetic sky in cosmology

Things We Want To Do With AI In The Future

- Develop methods that can learn from both encoded symbolic theory (e.g. QM/GR) and large-scale data so we can leverage the vast theoretical knowledge we have accumulated over hundreds of years
- Automate and accelerate discovery from planning, to conjecture, to experiment, to confirmation and analysis ⇒ end-to-end automated science
- Create an ability to use AI for generating new theories that address the problematical areas of existing theories

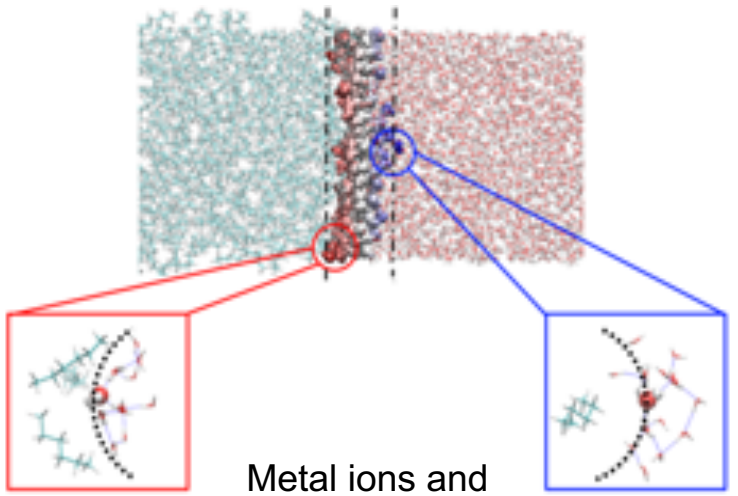
In Ten Years…

- **Learned Models Begin to Replace Data**
  - queryable, portable, pluggable, chainable, secure
- **Experimental Discovery Processes Dramatically Refactored**
  - models replace experiments, experiments improve models
- **Many Questions Pursued Semi-Autonomously at Scale**
  - searching for materials, molecules and pathways, new physics
- **Simulation and AI Approaches Merge**
  - deep integration of ML, numerical simulation and UQ
- **Theory Becomes Data for Next Generation AI**
  - AI begins to contribute to advancing theory
- **AI Becomes Common Part of Scientific Laboratory Activities**
  - Infuses scientific, engineering and operations

A Sampling of Science Opportunities

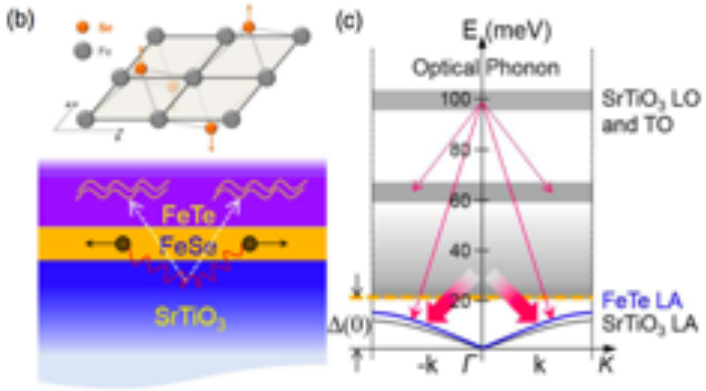
Materials and Chemistry

- Design of materials and molecules
- AI-guided synthesis
  - automated design of chemical pathways
  - mapping metastable phases
  - extracting mechanisms
- Predictive interfacial transport of ions and charge
- AI-accelerated ab Initio molecular dynamics
- Quantification of energy drivers for separations
- Describing multiscale charge, spin, lattice correlations
- Exploring energy landscapes in ultrafast, nonequilibrium, and driven systems and processes
- Inverse design, bandstructure engineering



Metal ions and complexes at liquid interfaces

Tian et al. PRL 2016



Nonequilibrium superconductivity

ML models achieve impressive results for many materials problems

Table 1 Materials informatics model results from the literature. The Pearson correlation coefficient R between predicted and actual property values is a common means of quantifying model performance. RMSE is root mean square error, MAE is mean absolute error, R^2 is the square of the Pearson correlation coefficient

Material class	Property	ML technique	CV type	Model performance metric	Ref.
Steel	Fatigue strength	Multivariate polynomial regression	Leave-one-out CV	R^2 = 0.9881	3
Organic small molecules	Norm of dipole moment	Graph convolutions	Overall 90% train/10% test, with reported test error averaged across 10 different models built on subsets of training data	MAE = 0.104 Debye (chemical accuracy target: 0.10 Debye)	4
Polymers	Electronic dielectric constant	Kernel ridge regression	85% train/15% test	R^2 = 0.96	16
Inorganic compounds	Formation energy	Rotation forest	32% train/68% test	R^2 = 0.93	5
Inorganic compounds	Vibrational free energy	Random forest or support vector machine	10 averaged k-fold CV runs, for k in [ref. 5 and 14]	R = 0.95	6
Inorganic compounds	Band gap	Support vector machine	100 averaged 75% train/25% test runs	G_0W_0 RMSE = 0.18 eV (DFT RMSE ~2 eV wrt expt.)	7

Meredig et al., DOI:10.1039/c8me00012c

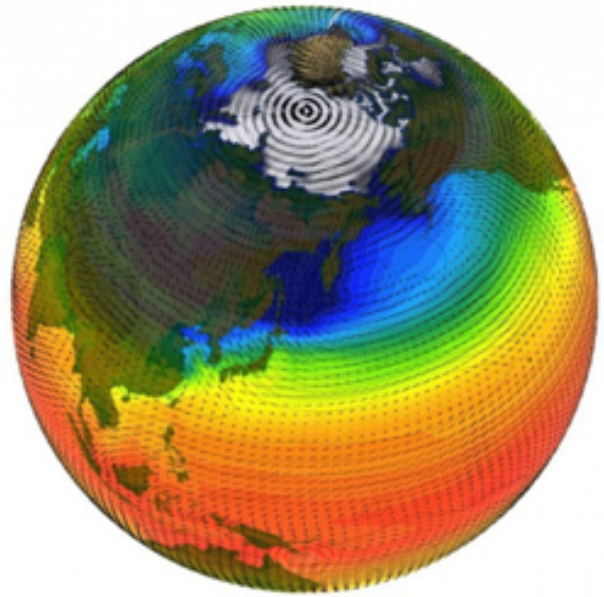
Implications for ML-guided materials design

Uncertainty quantification on top of ML models is crucial to evaluating candidates in new regions of design space

ML models are more useful as guides for **an iterative sequence of experiments**, as opposed to single-shot screening tools that can reliably evaluate an entire search space once and shortlist high-performing materials

Meredig et al., DOI:10.1039/c8me00012c

Climate and Biology



- Accelerated Climate Models (PDE/ML hybrids)
- Improved integration of remote sensing and ground truthing into Climate Models (cloud/precipitation, land cover/biogeochem, sea ice/calibration, etc.)
- Improvement in ARM data pipelines, automated model extraction from data, smart data fusion

- Vast applications in genomics and metagenomics (G $\Longrightarrow$ P)
- Automation of bioinformatics methods (improved productivity)
- Automating hypothesis formation in biology (causal analysis)
- Forward design of novel pathways, proteins, regulons, operons, organisms, etc. for secure biodesign
- Anomaly detection (discovery in sequencing, biosecurity, etc.)

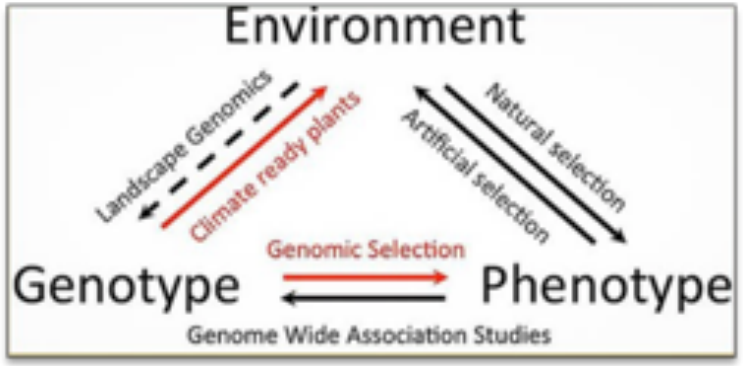


Fig. 1 KMR for drug representation learning and drug-drug interaction prediction. Red, green, blue and yellow matrices denote pharmacological feature representations, drug textual description feature representations, drug class feature representations, and final knowledge-oriented drug representations, separately

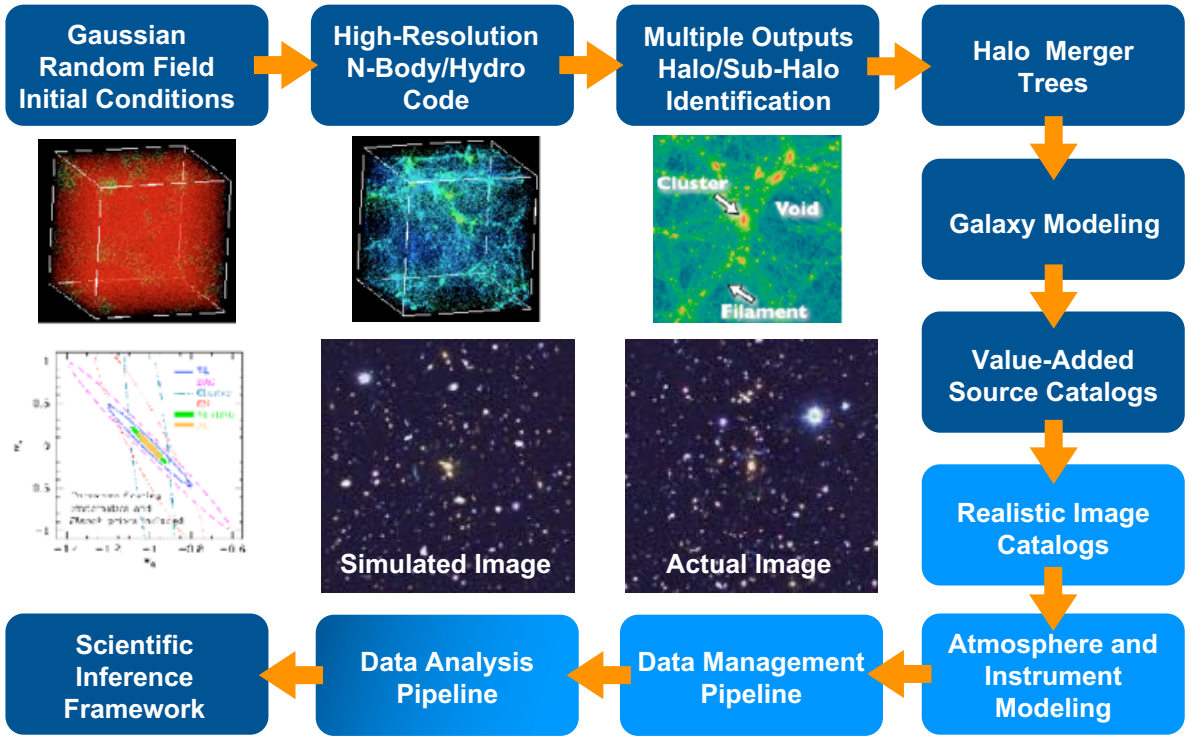
High Energy Physics

Energy/Intensity Frontier:

- Search for Beyond the Standard Model (BSM) physics through AI-driven anomaly detection
- AI-reduced uncertainties to enable precision electroweak measurements for BSM clues
- Generative Adversarial Networks (GANs) for large-scale Large Hadron Collider detector simulation

Cosmic Frontier – AI in end-to-end application:

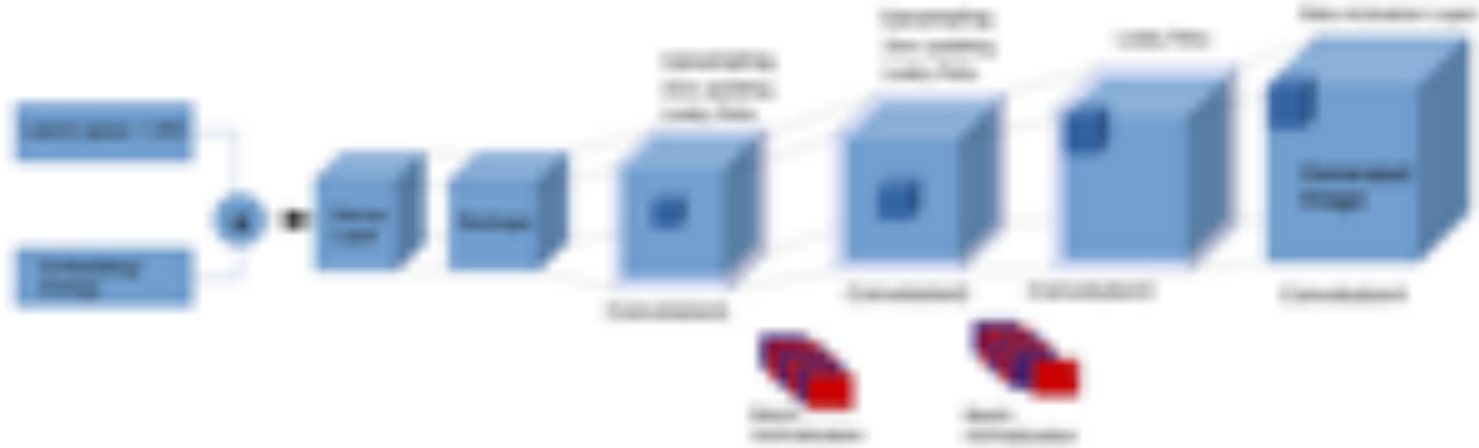
- Precision Cosmic Microwave Background emulation – AI simulation speed-up of a factor of 1000
- Search for strong lensing of galactic sources for precision cosmology measurements using AI classification, regression, and GANs for image generations
- AI-based Photometric Redshift Estimation
- Combination of AI methods to enable searches for hidden space variables

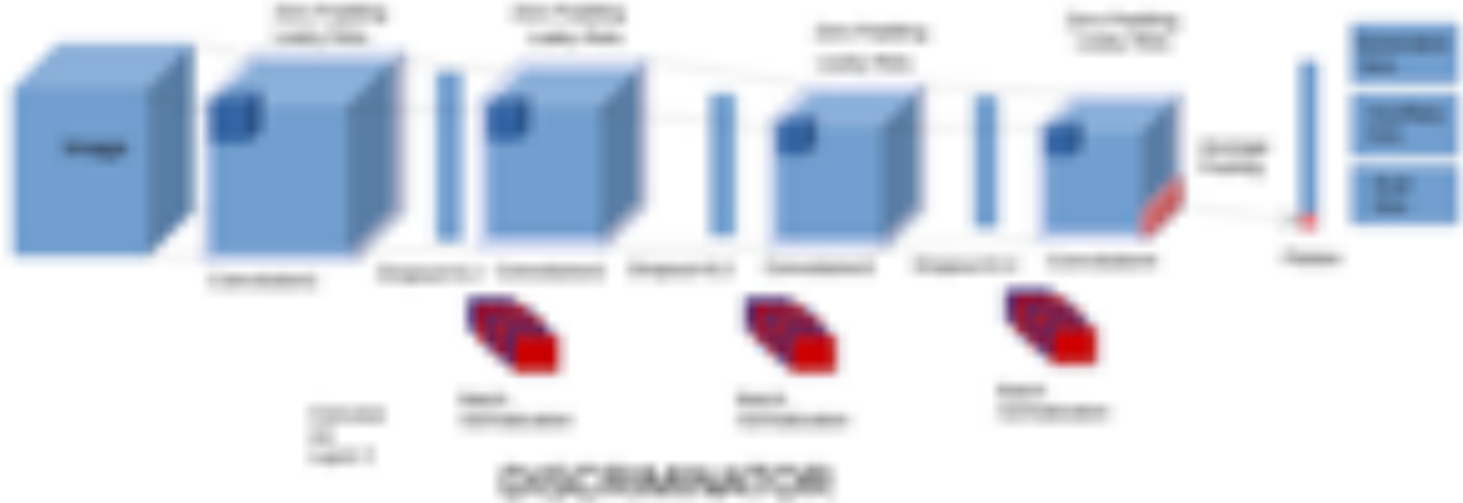


AI applications in an "end-to-end" Cosmic Frontier application: 1) GANs for image emulation, 2) GP and DL-based emulators for summary statistics, 3) CNN-based image classification, 4) AI-based photometric reshift estimation, 5) Likelihood-free methods for inference *[Work performed under the Argonne-led SciDAC-4 project: "Inference and Machine Learning at Extreme Scales"]*

3D convolutional GAN

- Similar discriminator and generator models
- 3D convolutions (keep X,Y symmetry)
- Tested several tips&tricks from literature[*]
- Some helpful (no batch normalisation in the last step, LeakyRelu, no hidden dense layers, no pooling layers)
- RMSProp optimiser for both networks
- Batch training



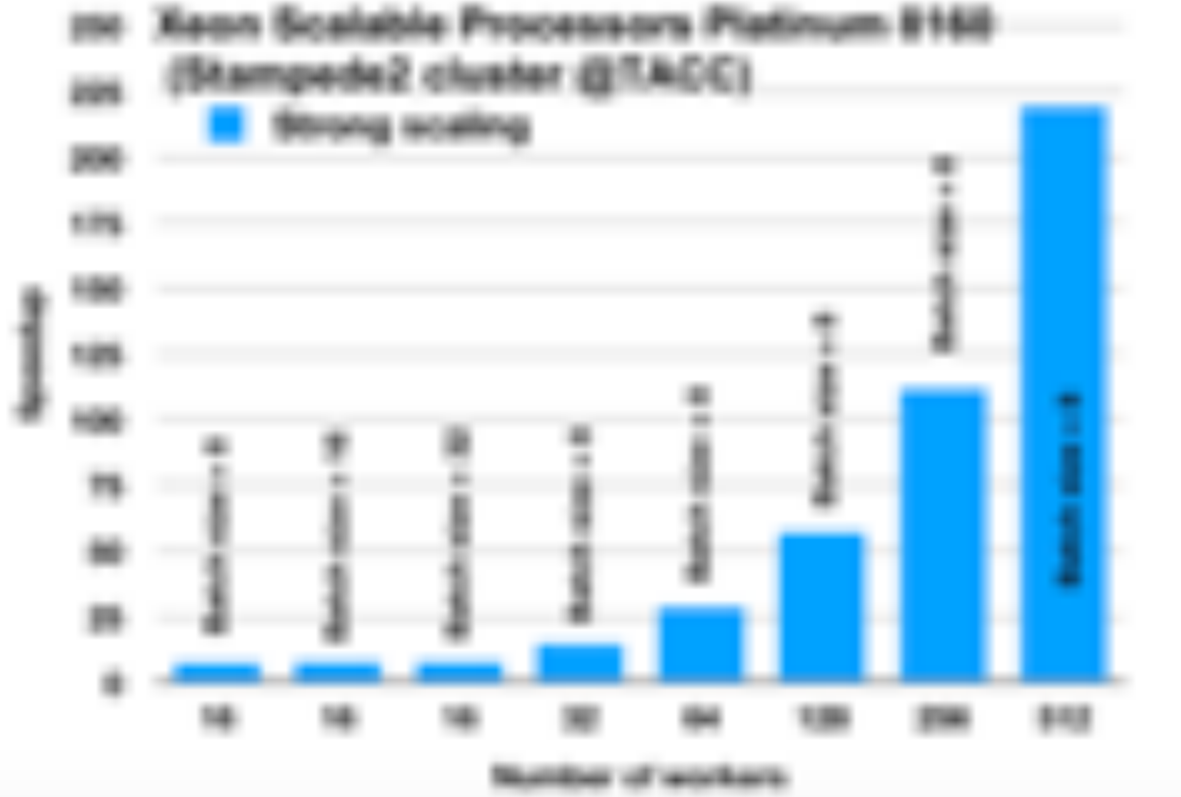


[*] https://github.com/soumith/ganhacks

Computing performance

Time to create an electron shower		
Method	Machine	Time/Shower (msec)
Full Simulation (geant4)	**Intel Xeon Platinum 8180**	**17000**
3d GAN (batch size 128)	**Intel Xeon Platinum 8180**	**7**
3d GAN (batchsize 128)	GeForce GTX 1080	0.04

- Inference: speedup factor > 2500



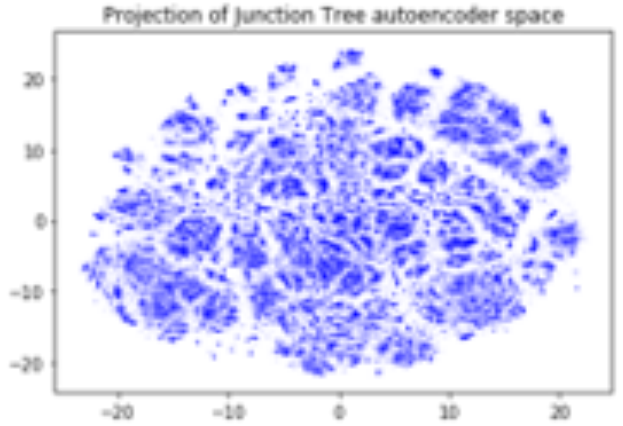
- Training (speedup vs. performance)
  - 45 minepoch on Tesla P100
  - Introduce data parallel training based on MPI
  - Test several libraries
  - More in the MS04 mini-symposium

Connecting HPC and AI

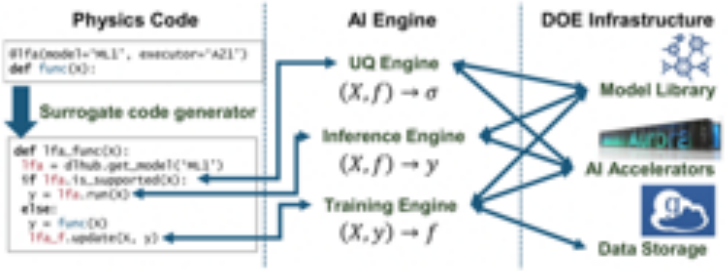
In addition to partnerships in AI applications, there are considerable opportunities in foundational methods development, software and software infrastructure for AI workflows and advanced hardware architectures for AI, below we highlight some ideas in the HPC + AI space

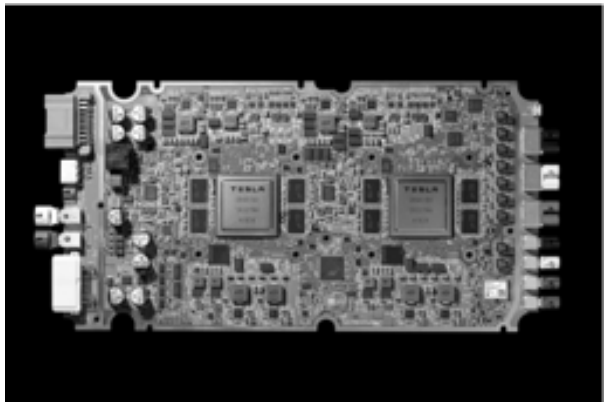
- Steering of simulations
- Embedding simulation into ML methods
- Customized computational kernels
- Tuning applications parameters
- Generative models to compare with simulation
- Student (AI) Teacher (Sim) models ⟹learned functions
- Guided search through parameter spaces
- Hybrid architectures HPC + Neuromorphic
- Many, many more

Generative Models



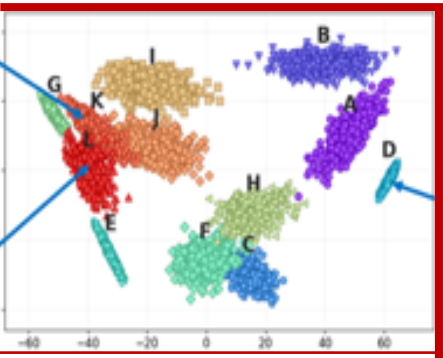
Learned Function Accelerators



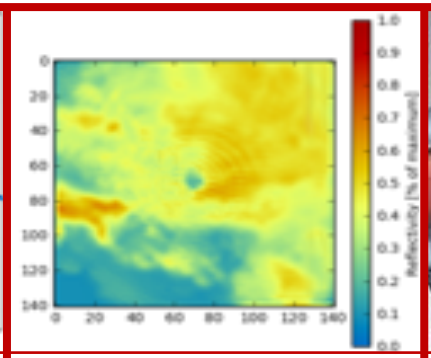


AI Accelerators

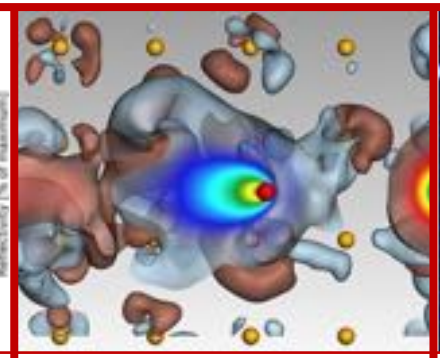
AI at Argonne: Broad Span of Scientific Targets



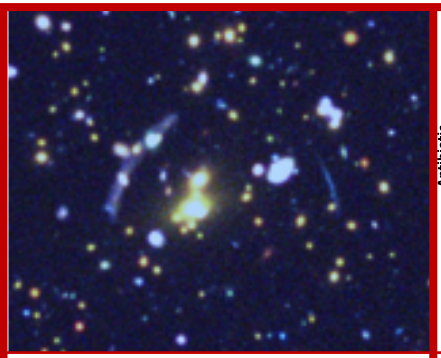
Reduced order modeling of laser sintering



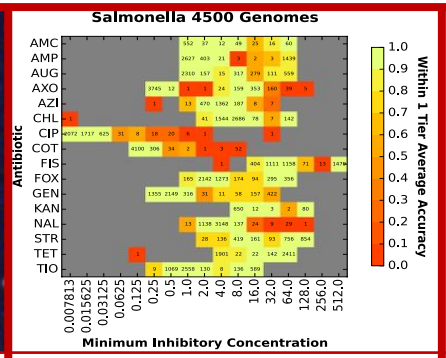
Nowcasting with convolutional LSTMs



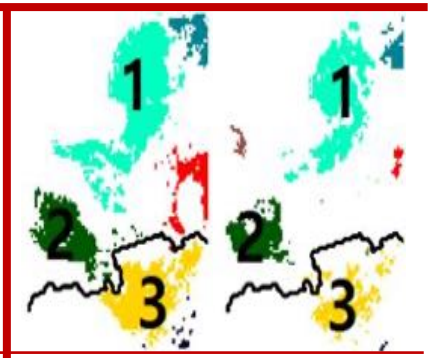
Prediction of radiation stopping power



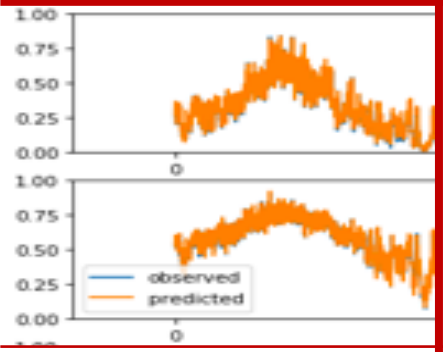
Strong and weak lensing in sky survey data



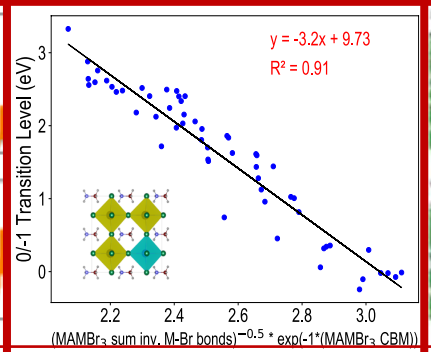
Prediction of antimicrobial resistance phenotypes



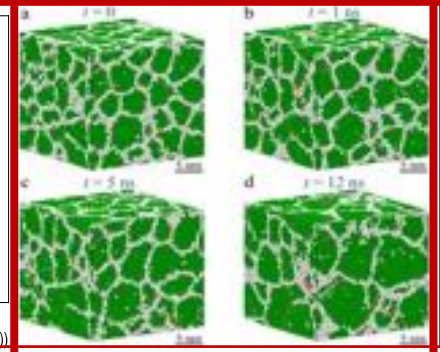
Identification and tracking of storms



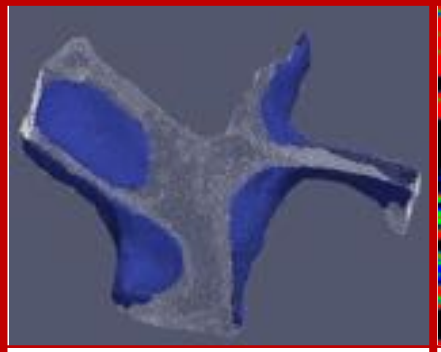
Efficient climate model emulators



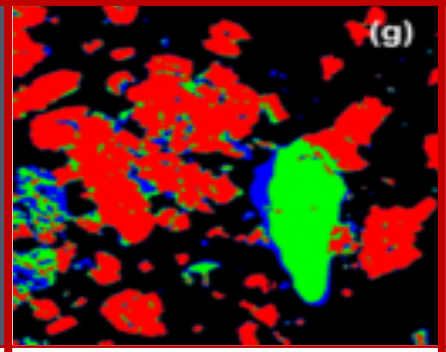
Defect-level prediction in semiconductors



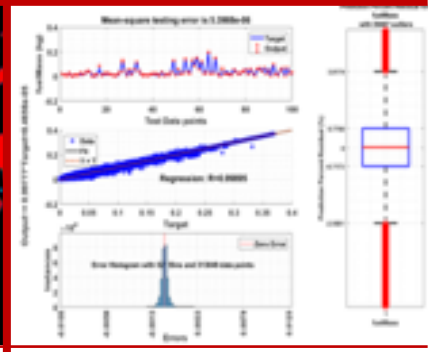
Structure-property-process triangle in additive manufact.



Parameter extraction in atom probe tomography

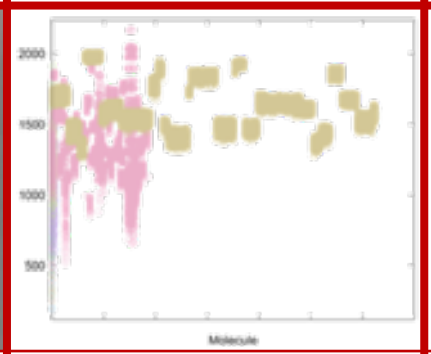


Learning for dynamic sampling in spectroscopy

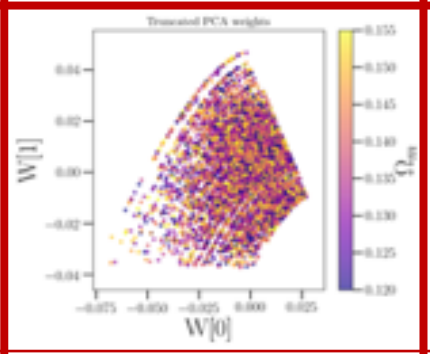


Vehicle energy consumption prediction

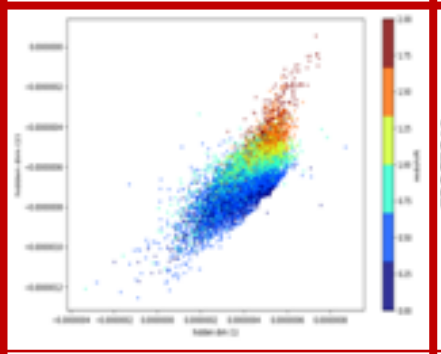
Flying object detector for edge deployment



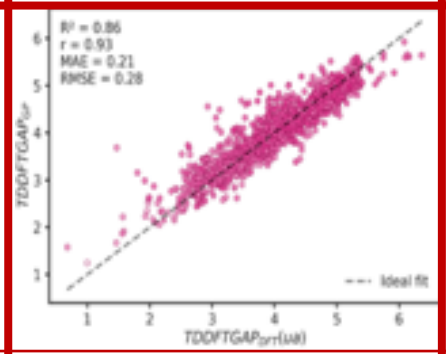
Discovery of new energy storage materials



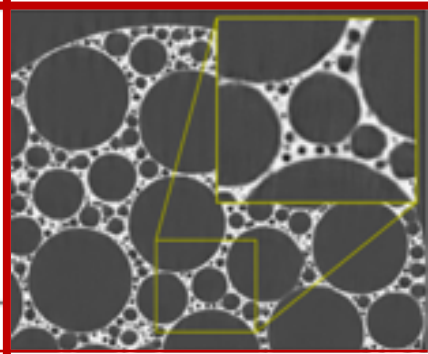
Cosmic Microwave Background emulation



Photometric red shift estimation



New materials for efficient solar cells



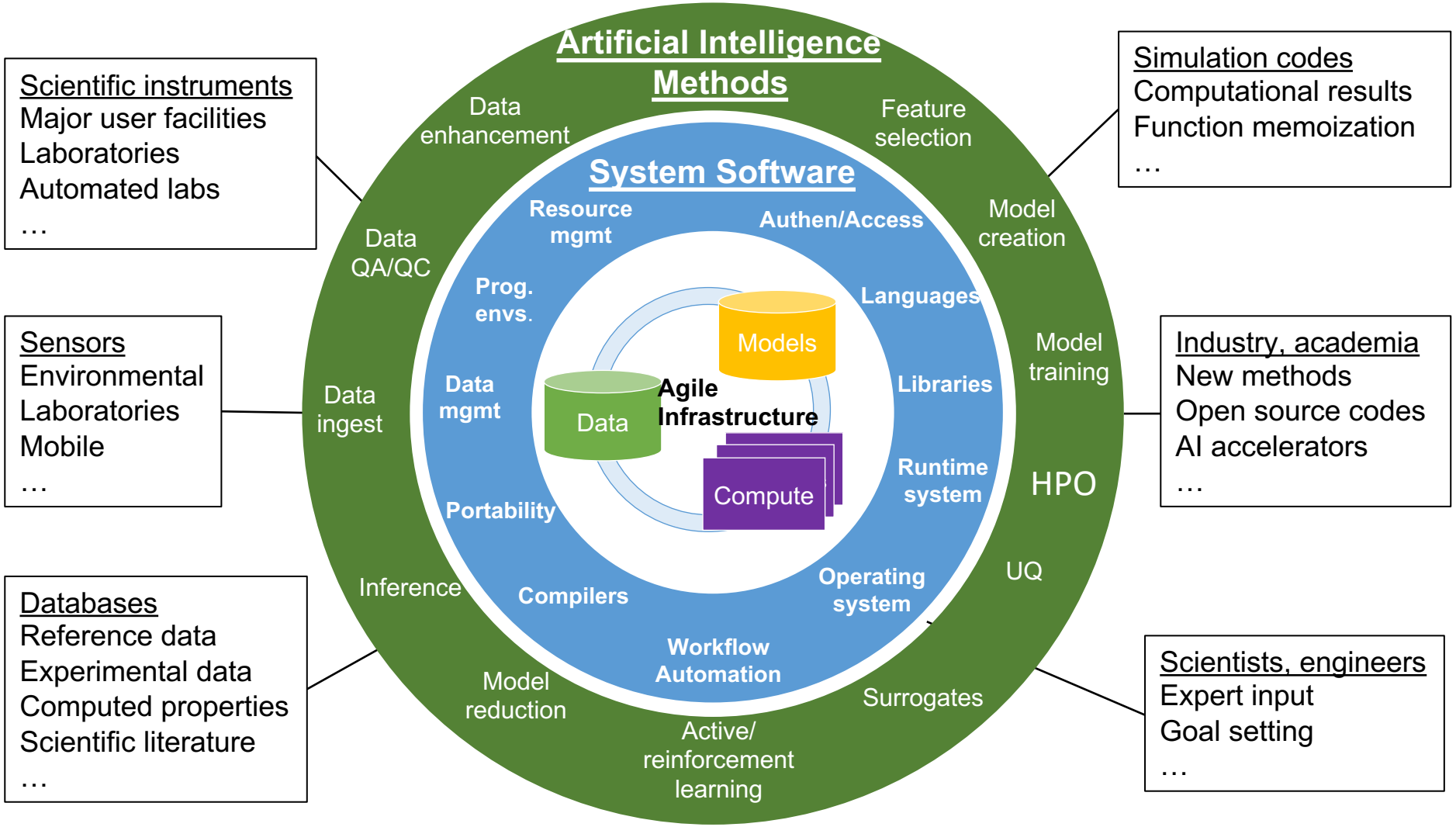
Enhancement of noisy tomographic images

Building the AI Environment for Science

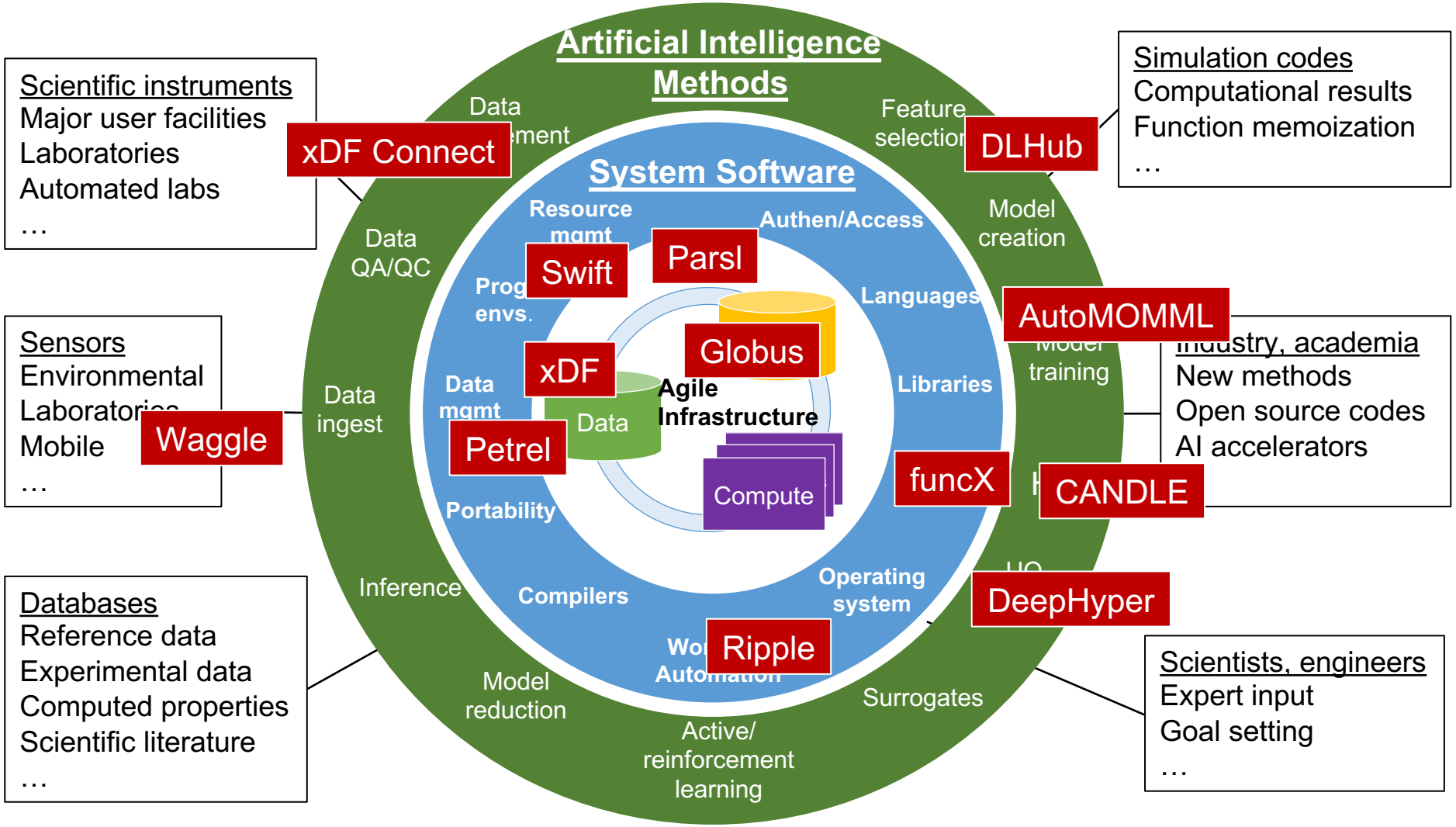
AI for Science Requires New Research and Infrastructure

Applications	AI applications across science and engineering. Transformative approaches to simulation and experimental science.
Learning systems	AI software. Software infrastructure for managing data, models, workflows etc., and for delivering AI capabilities to 1,000s of scientists and engineers.
Foundations	Mathematics, algorithms; general AI, reinforcement learning, uncertainty quantification, explainability, etc.
Hardware	Advanced hardware to support AI. Evaluation of new architectures and systems; exploration of neuromorphic and quantum as long term accelerators for AI.

Infrastructure for AI-enabled Science



Infrastructure for AI-enabled Science



DLHub: Organizing and Serving Models

https://www.dlhub.org

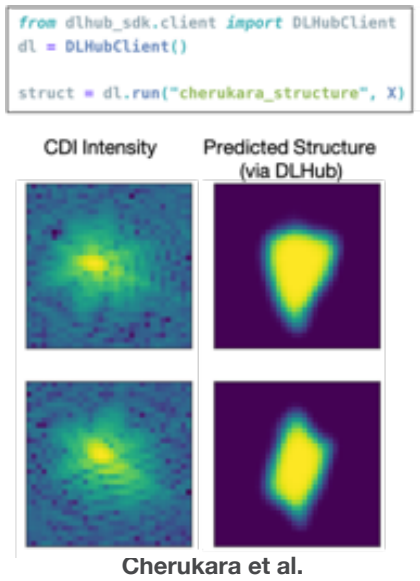
DLHub
Data and Learning Hub for Science
https://www.dlhub.org

- Collect, publish, categorize models
- Serve models via API with access controls to simplify sharing, consumption, and access
- Leverage ALCF resources and prepare for Exascale ML
- Deploy and scale automatically
- Provide citable DOI for reproducible science

Argonne Advanced Computing LDRD

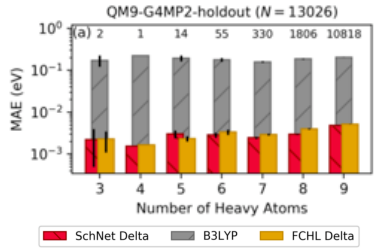
Models and Processing Logic as a Service

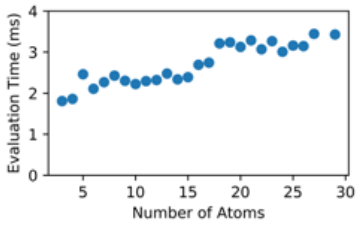
X-Ray Science



Cherukara et al.

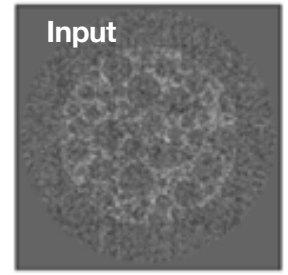
Energy Storage

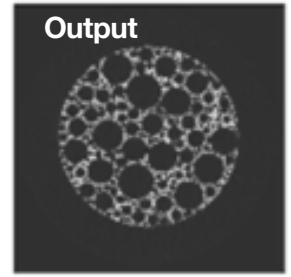




Ward et al.

Tomography





TomoGAN: Liu et al.

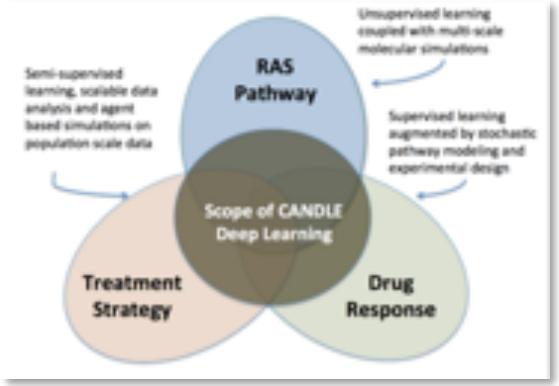
CANDLE: Exascale Deep Learning Tools

Deep Learning Needs Exascale

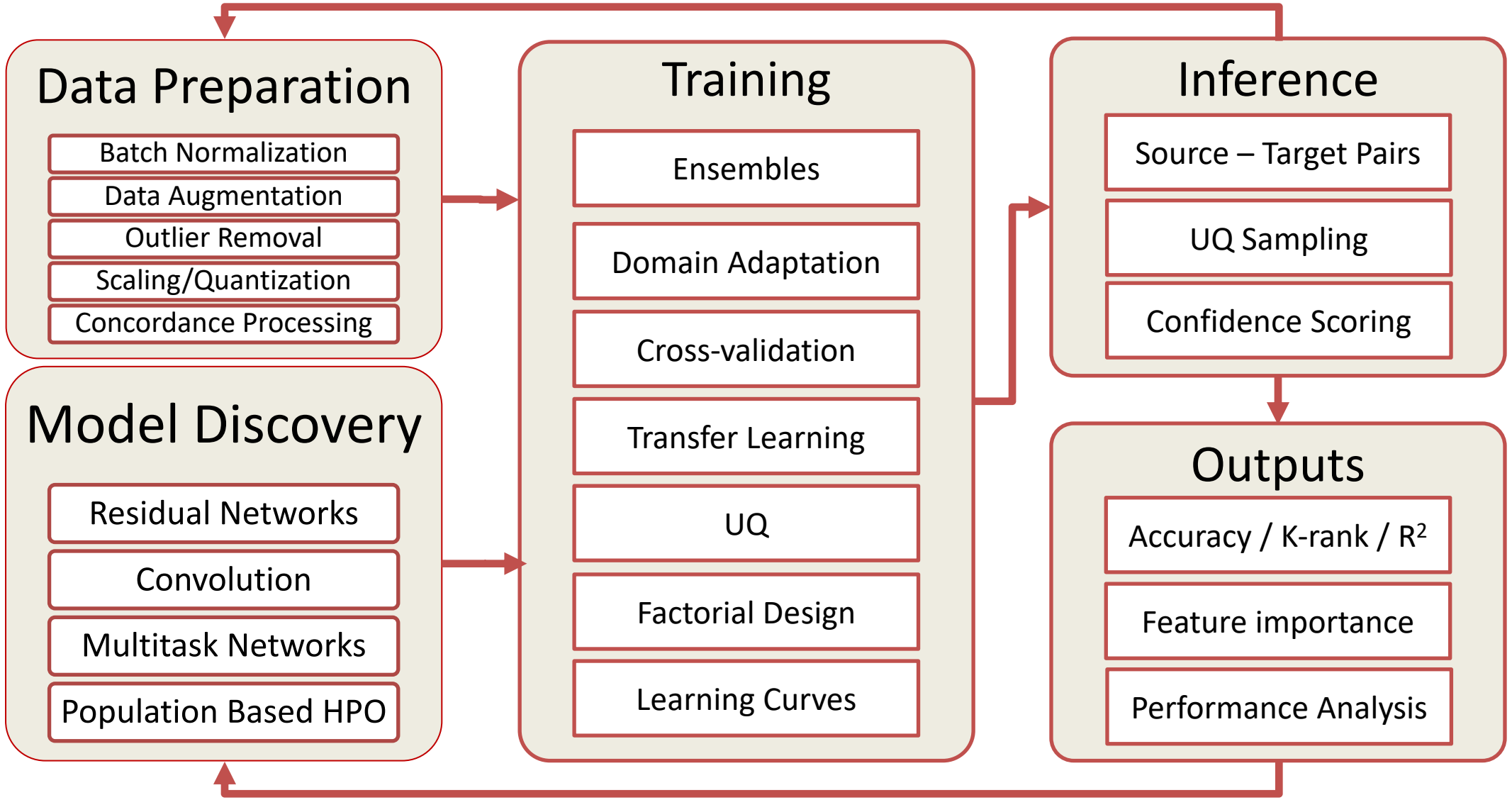
- Automated model discovery
- Hyper parameter optimization
- Uncertainty quantification
- Flexible ensembles
- Cross-Study model transfer
- Data augmentation
- Synthetic data generation
- Reinforcement learning

- **CANDLE Python Library** – make it easy to run on DOE Big Machines, scale for HPO, UQ, Ensembles, Data Management, Logging, Analysis
- **CANDLE Benchmarks** – exemplar codes/models and data representing the three primary challenge problems
- **Runtime Software** – Supervisor, Reporters, Data Management, Run Data Base
- **Tutorials** – Well documented examples for engaging the community
- **Contributed Codes** – Examples outside of Cancer, including Climate Research, Materials Science, Imaging, Brain Injury
- **Frameworks** – Leverage of TensorFlow, Keras, Horovod, PyTorch, etc.
- **LL Libraries** – CuDNN, MKL, etc. (tuned to DOE machines)

https://github.com/ECP-CANDLE

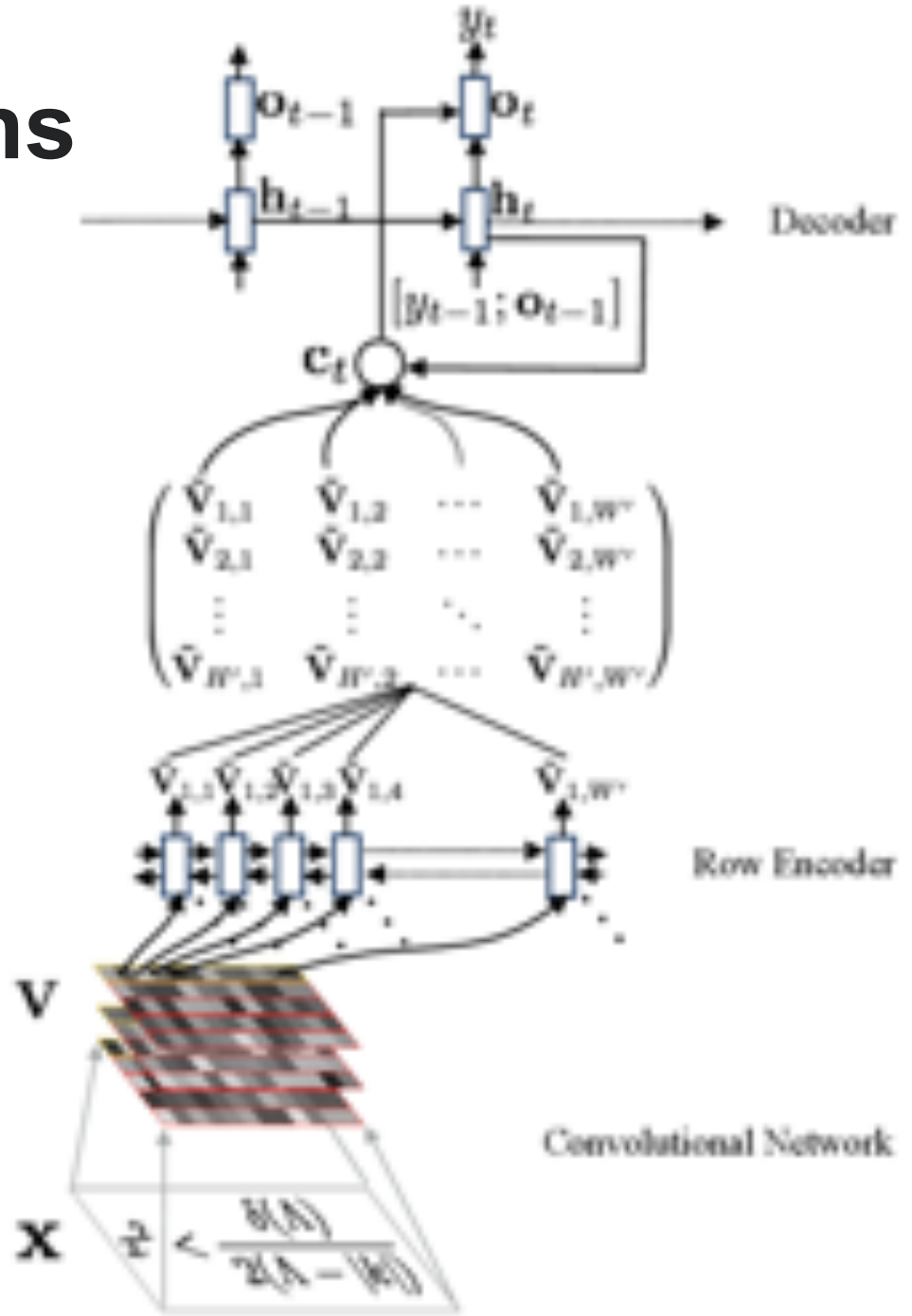


Scope of CANDLE workflows



Future Directions in Foundations

- Leverage DOE expertise in automatic differentiation, symbolic computing and optimization to ensure that machine learning for science is forward looking, methods are robust and models interpretable
- Many facets relevant to science
  - Integration of symbolic computing with machine learning
  - Prediction and inference of spatio-temporal processes
  - Derivatives for training, sensitivity analysis, optimization, and UQ
  - Rapid data analysis to reduce volume or identify features of interest
  - Variety of new approaches to inference and UQ
  - Identify and account for uncertainty in data sources and computations



Methods Innovation, one page agenda ☺

- Data efficient learning. “Low data”. One shot, few shot learning
- Improved neural architecture search. Lottery tickets and sparsity.
- Online learning and incremental training. Active learning.
- Representation learning in novel “scientific” spaces
- UQ and confidence estimates. Interpretability.
- Integration of symbolic computing and deep learning. Synthesis learning.
- Beyond NLP and CV towards concepts directly needed by science.
- Generative methods in scientific and engineering domains.
- Inverse methods and systems that input data and output rules.


We are starting out in a good place

Aurora: HPC and AI

Argonne NATIONAL LABORATORY

>> Exaops/s for AI

intel

CRAY

Architecture supports three types of computing

- Large-scale Simulation (PDEs, traditional HPC)
- Data Intensive Applications (scalable science pipelines)
- Deep Learning and Emerging Science AI (training and inferencing)

DEPARTMENT OF ENERGY · UNITED STATES OF AMERICA

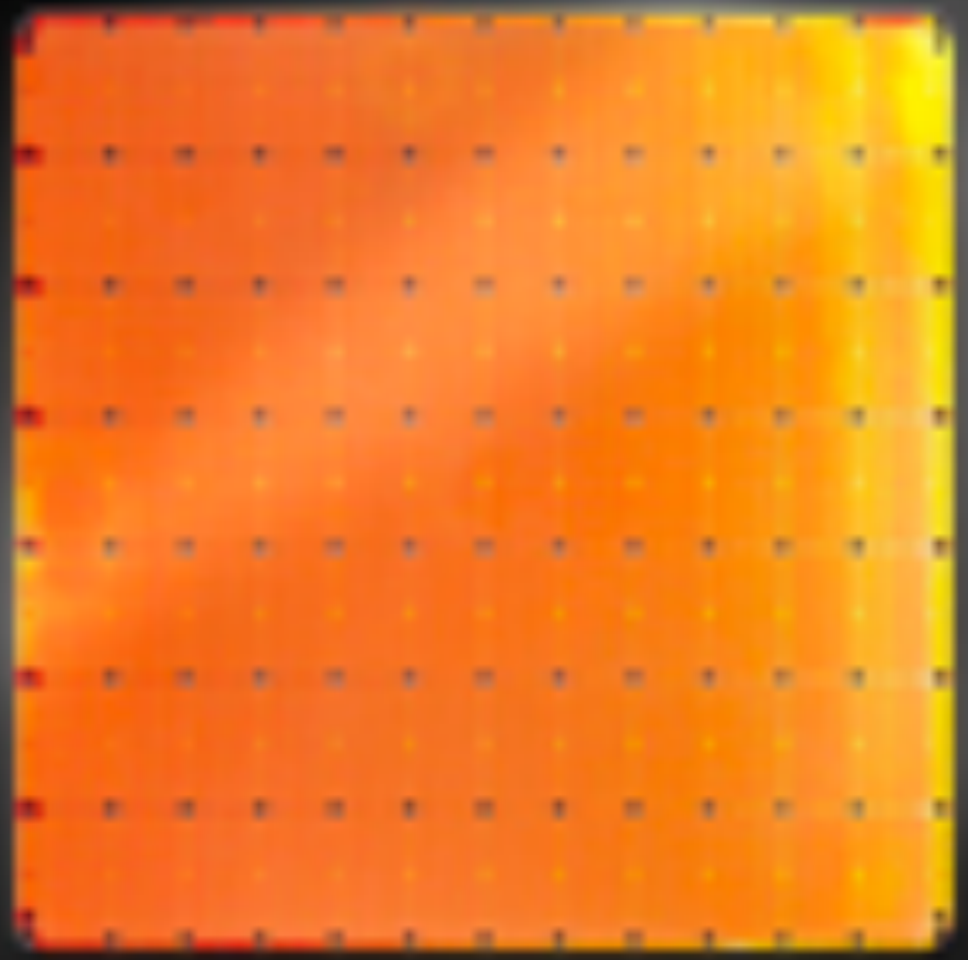
Specialized hardware is emerging that will be 10x – 100x the performance of general purpose CPU and GPU designs for AI

US VCs investing >$4B in startups for AI acceleration

Which platforms will be good for science?

AI Chip Landscape

Cerebras Wafer Scale Engine



Cerebras WSE

1.2 Trillion Transistors

46,225 mm² Silicon

Largest GPU

21.1 Billion Transistors

815 mm² Silicon

AI Accelerator Testbed

Engaging the community to understand and improve specialized AI hardware for science

Dozens of proposed AI accelerators promise
10x - 1000x acceleration for AI workloads. AI testbed will:

1. Provide an **open and unbiased environment** for evaluation of AI accelerator technologies
2. **Disseminate information** about use cases, software, performance on test problems
3. **Support collaborations** with AI technology developers, academics, commercial AI, DOE labs

IC Vendors	Intel, Qualcomm, Nvidia, Samsung, AMD, Xilinx, IBM, STMicroelectronics, NXP, Marvell, MediaTek, HiSilicon, Rockchip	13
Tech Giants & HPC Vendors	Google, Amazon_AWS, Microsoft, Apple, Aliyun, Alibaba Group, Tencent Cloud, Baidu, Baidu Cloud, HUAWEI Cloud, Fujitsu, Nokia, Facebook, HPE, Tesla	12
IP Vendors	ARM, Synopsys, Imagination, CEVA, Cadence, VeriSilicon, Videantis	7
Startups in China	Cambricon, Horizon Robotics, Bitmain, Chipintelli, Thinkforce, Unisound, AISpeech, Rokid, NextVPU, Canaan, EnFlame, Eeasy Tech	12
Startups Worldwide	Cerebras, Wave Computing, Graphcore, PEZY, Tenstorrent, ThinCI, Koniku, Adapteva, Knowm, Mythic, Kalray, BrainChip, Almotive, DeepScale, Leepmind, KrtKl, NovuMind, REM, TERADEEP, DEEP VISION, Groq, KAIST DNPU, Kneron, Esperanto Technologies, Cyrfalcon Technology, SambaNova Systems, GreenWaves Technology, Lightelligence, Lightmatter, ThinkSilicon, Innogrit, Kortiq, Halio, Tachyum, AlphaICs, Syntiant, Habana, aiCTX, Flex Logix, Preferred Network, Cornami, Analflash, Optalysys, Eta Compute	44

Staged evaluation enables identification of most promising systems for science

Device

Subrack

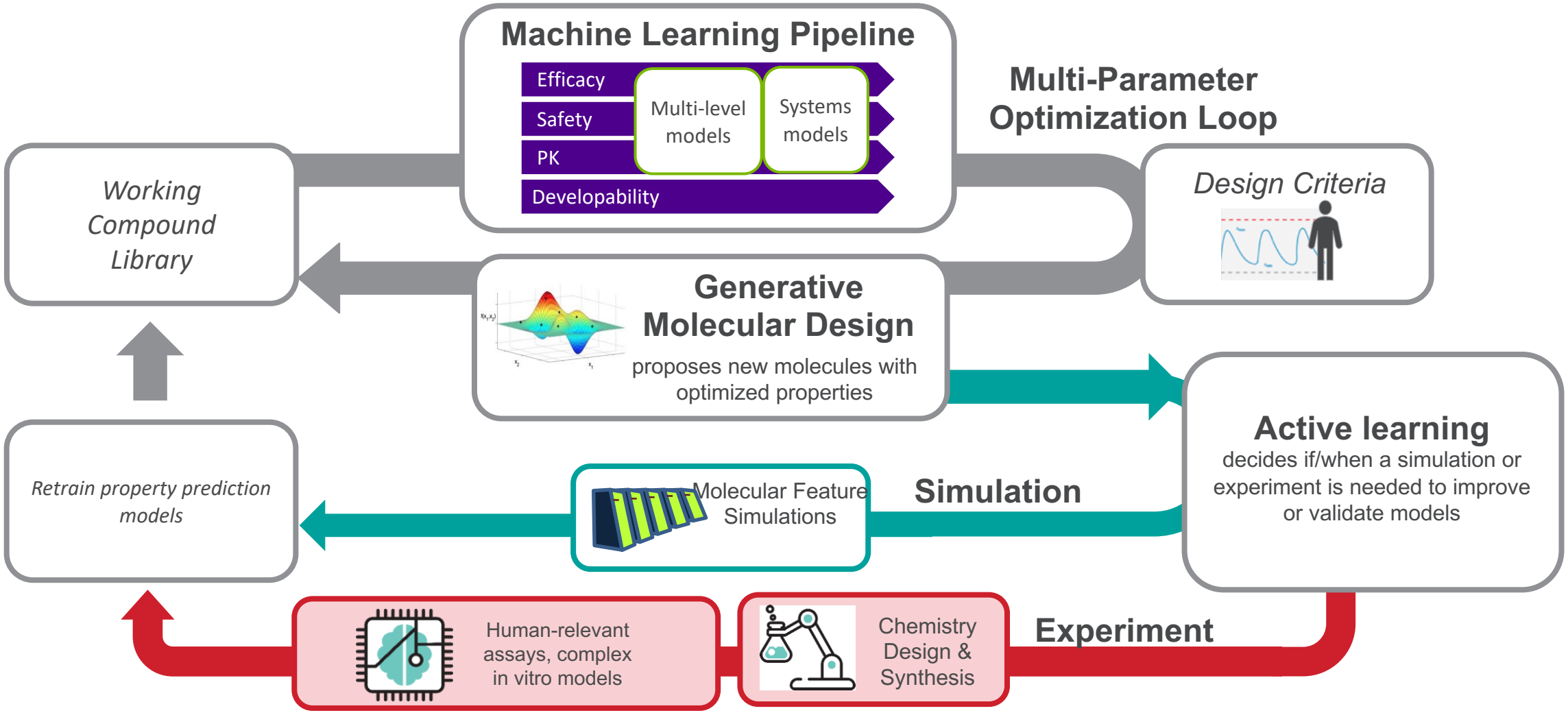
Rack

https://github.com/basicmi/AI-Chip


AI Driven Experimental Science

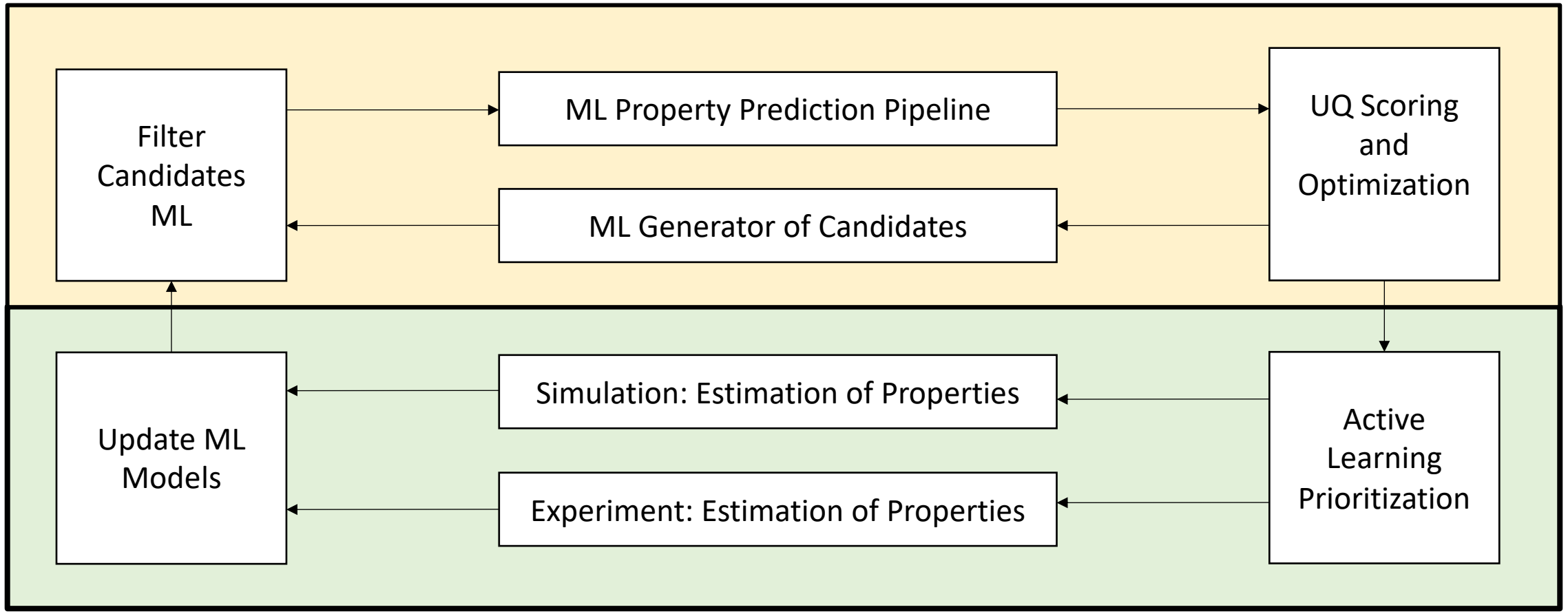
The ATOM Platform

Active Learning Drug Discovery Framework



Jim Brase (LLNL) and the ATOM Consortium

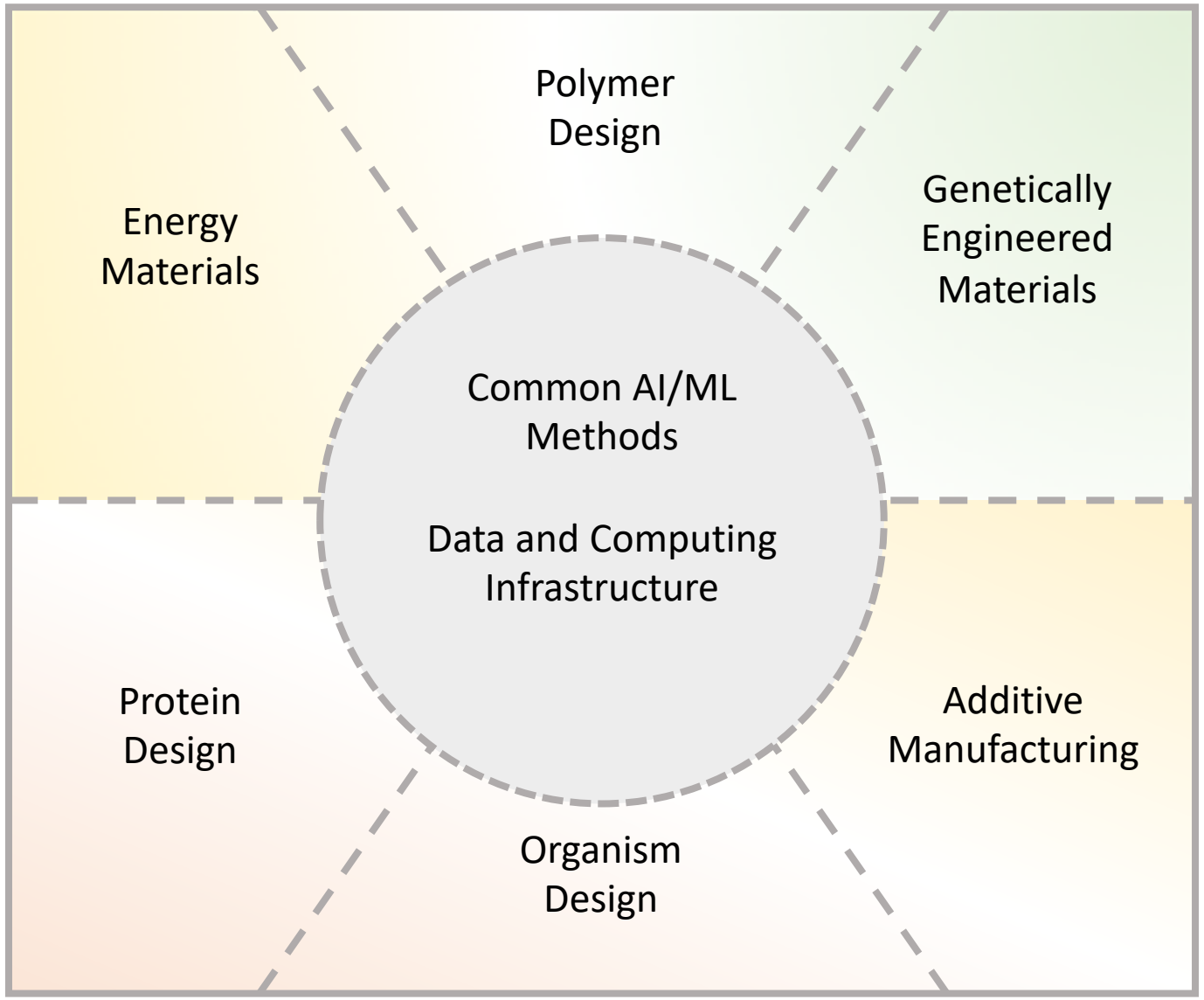
Layered workflow combining AI, HPC and HTS



Pure ML “constant time” (fast loop)

Mixed/Variable time (slow loop)

AI Driven Autonomous Laboratory Cluster



Overall Lessons Learned

- **Aggregation, integration, normalization and curation of data is critical**
  - Assay methodology and interpretation of responses is fundamental to understanding modeling performance
- **Impact of data scale and data quality are deeply intertwined**
  - Study design matters
  - Variety of approaches are needed
- **Model target use matters on validation strategy and tuning**
  - We can optimize model performance for different use cases
- **Large-scale computing enables us to dive deep on some questions**
  - Impact is more confidence in the modeling choices, and deep understanding of alternatives
- **Frequent meetings and interactions (Hackathons) are critical**
  - Bridging the language and scientific culture differences requires time