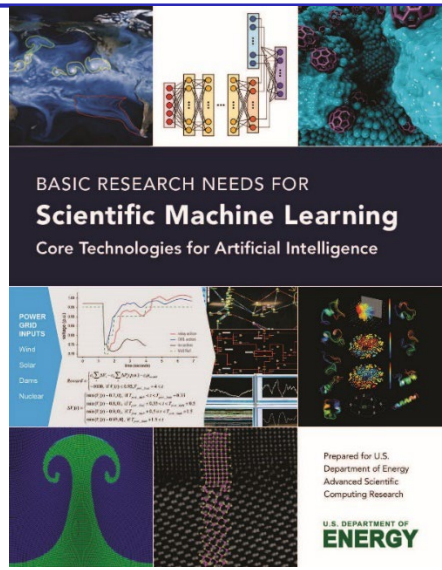
HPC Challenges for New Extreme Scale Applications
March 6-7, 2023

ML/AI Research Directions within the US Department of Energy

Osni Marques
Lawrence Berkeley National Laboratory
oamarques@lbl.gov

Research Directions

- Foundational research themes of Scientific Machine Learning
- Opportunities for Scientific Machine Learning impact



SciML Foundations

Machine Learning for Advanced Scientific Computing Research

Domain-aware leveraging & respecting scientific domain knowledge	physical principles & symmetries physics-informed priors structure-exploiting models ⋮
Interpretable explainable & understandable results	model selection exploiting structure in high-dim data uncertainty quantification + ML ⋮
Robust stable, well-posed & reliable formulations	probabilistic modeling in ML quantifying well-posedness reliable hyperparameter estimation ⋮

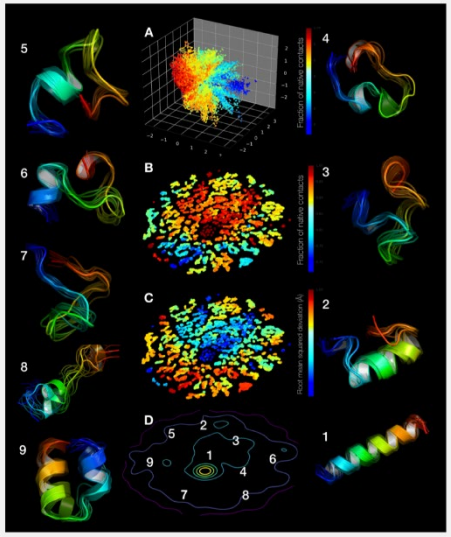
SciML Capabilities

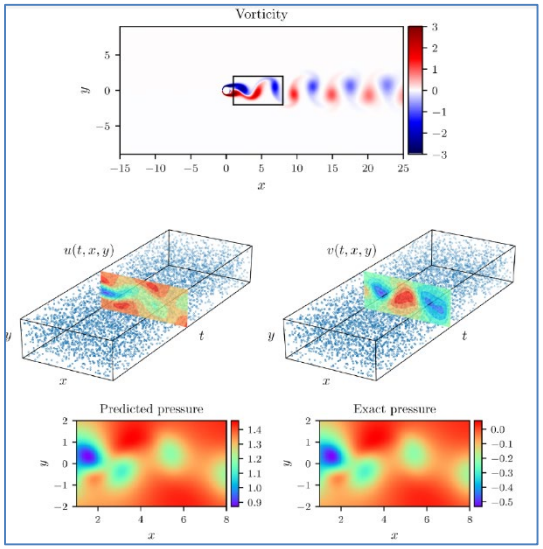
Machine Learning for Advanced Scientific Computing Research

Data-intensive scientific inference & data analysis	ML methods for multimodal data in situ data analysis with ML ML to optimally guide data acquisition ⋮
ML-enhanced modeling & sim ML-hybrid algorithms and models for better scientific computing tools	ML-enabled adaptive algorithms ML parameter tuning ML-based multiscale surrogate models ⋮
Intelligent automation & decision support automated decision support, adaptivity, resilience, control	exploration of decision space with ML ML-based resource mgt & control optimal decisions for complex systems ⋮

Sample of Applications of Interest

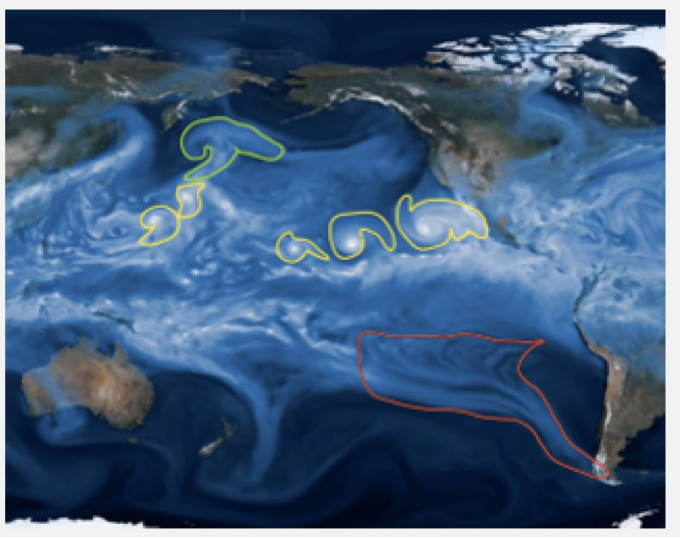
protein folding

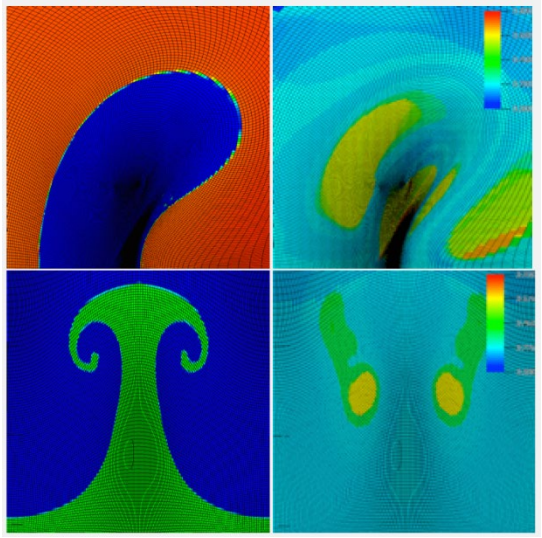




fluid dynamic simulations

extreme whether patterns





mesh relaxation

Research Thrust: Domain Awareness

- Incorporation of domain knowledge into unsupervised SciML and model feature selection
  - Features should be representative, interpretable, and generalizable
- Incorporation of domain knowledge into supervised SciML
  - Hard constraints: imposition of constraints that cannot be violated (enforced during training, projection into the constrained region)
  - Soft constraints: modify the objective function constraints used in training
  - Model form (e.g. symmetries and scaling in kernel approaches)
- Modeling and representing domain knowledge in SciML
  - modeling languages and frameworks that facilitate the inclusion of domain knowledge into the training process

Research Thrust: Interpretability

- Exploring high-dimensional complex data
  - methods that provide users with insights into data characteristics beyond traditional statistical indicators
- Exploring and understanding SciML models
  - rationalize or explain the relationship between the input, operation, and output
  - decision process for interpretable human-meaningful and human-manageable steps
- Expressing differences between SciML models, inputs and results
  - assist in the characterization of complex data sets
  - comparison of models at large scales

Research Thrust: Robustness

- Implementing reproducibility in SciML
  - an independent research group should be able to replicate the findings of another
  - “less than 50% of academic research by drug companies is reproducible”
- Conditions for “well-posedness”
  - models and algorithms that are insensitive to perturbations
- Assessing the robustness, performance and quality of SciML
  - decision (classification) or a prediction: traditional measures of acceptance are often heuristic
  - *a priori* and *a posteriori* error estimates would be transformative
  - algorithms that have proven convergence rates

Research Thrust: Data-Intensive SciML

- Extracting structures from high-dimensional data and complex models
  - improved methods for discovering sparsity and low-rank structure
  - learning underlying geometry beyond PCA
  - approaches for data compression or dimension reduction
- Efficiently sampling complex and high-dimensional spaces
  - improved methods for sampling, optimization, and integration in high-dimensional spaces
  - methods to allow the use of models at varying degrees of fidelity
- Achieving robustness in noisy and complex data
  - maximize the amount of learning from the available scarce data
  - data collection or generation procedures (observation/computation) to minimize the amount of data required and the associated costs

Research Thrust: Enhanced Modeling and Simulation

- Enabling adaptive scientific computing
  - improve the performance and throughput of numerical simulations through ML (e.g. by changing solver parameters)
  - ML to help in the choice of data layouts and architecture-aware algorithm implementations
- DOE's Scientific Computing help to SciML
  - expertise in scalable numerical algorithms
  - inner loop of the SciML training process (optimization algorithms, linear algebra solvers) at large scale
  - co-design of new and adaptation of existing SciML algorithms for different computer architectures

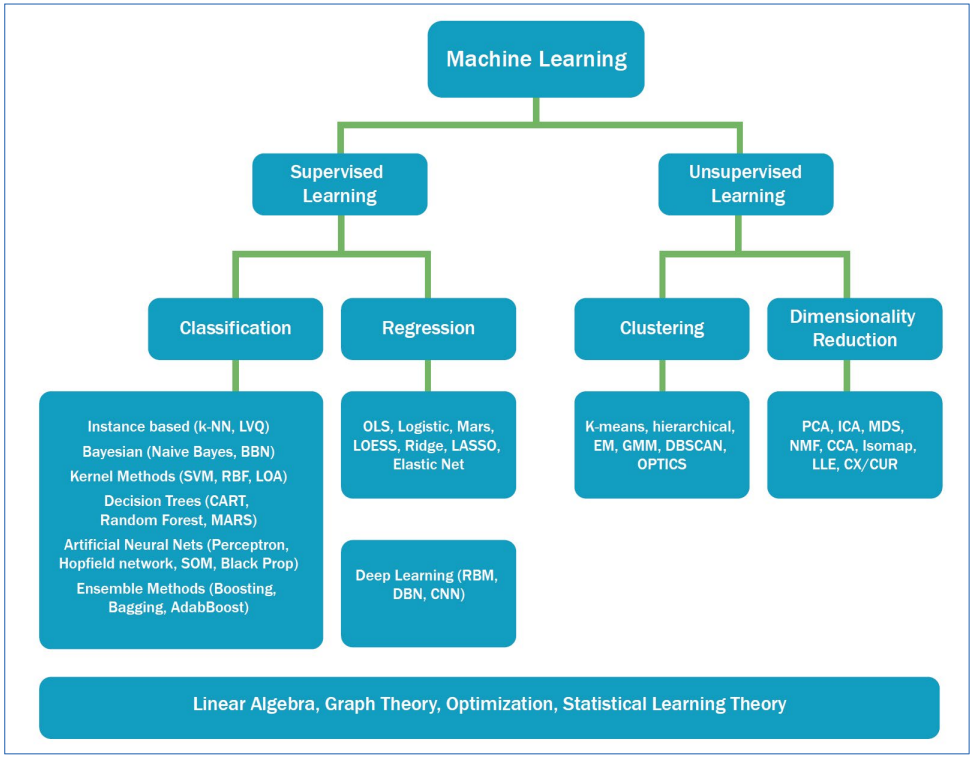
Research Thrust: Intelligent Automation and Decision Support

- ML to intelligently guide data acquisition
  - ML methods for optimal data acquisition strategies for applications of optimization, UQ, and sensitivity analysis
- ML to improve outcomes from science facilities
  - real-time monitoring of experiments (x-ray light sources, neutron scattering, magnetic fusion facilities, particle accelerators)
  - ML-based systems to track real-time telemetry data and predict failures of computational nodes

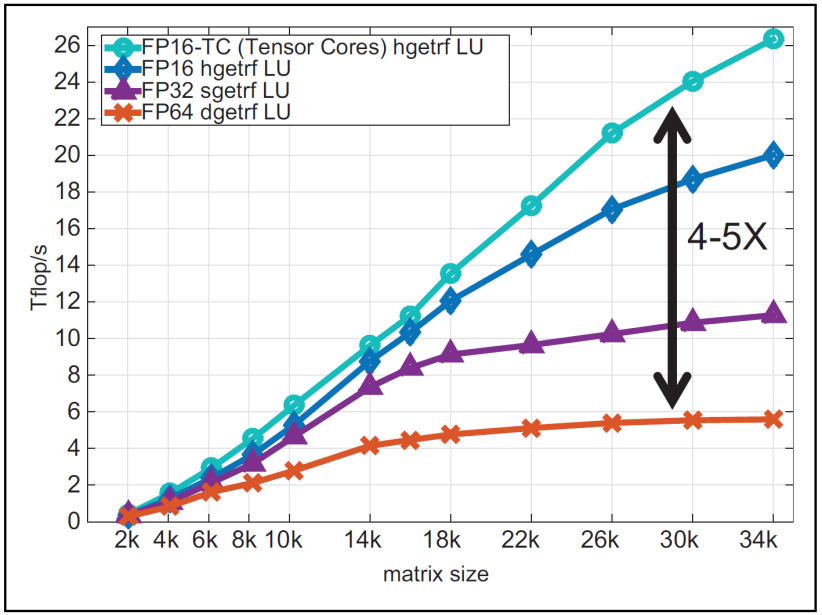
Numerical Calculations in SciML

- Matrix computations
- Numerical algorithms at large scale
  - Linear solvers
  - Optimization algorithms
  - Eigenvalue problems
  - Co-design
- Multi-precision algorithms

Simplified taxonomy, from Machine Learning and Understanding for Intelligent Extreme Scale Scientific Computing and Discovery Report, 2015



Multi-precision Computations



- Abdelfattah et al., *A survey of numerical linear algebra methods utilizing mixed-precision*, IJHPCA, Vol. 35, 2021.
- Anzt and Luszczek, *Accelerating Numerical Software Libraries with Multi-Precision Algorithms*, May 30, 2020, https://youtu.be/uG1L2slCqTs
- Anzt, *Then and Now – Growing as a child of ECP*, plenary talk at ECPAM 2023

Multi-precision: take-aways (1/2)

(from Anzt and Luszczek webinar)

- Performance of compute-bound algorithms depends on format support of hardware
- Performance of memory-bound algorithms is hardware-independent and scales with the inverse of format complexity
- Relative residual accuracy = $\varepsilon\kappa$ (unit round-off * condition number)
- For ill-conditioned problems, we need high precision to provide high accuracy results.
- Only if the problem is well-conditioned, and a low-accuracy solution is acceptable, we can use a low precision format throughout the complete solution process
- Templating the precision format allows to quickly switch between formats
  - C++ very powerful in this respect
  - Use production-ready libraries templating the precision: Ginkgo, Kokkos Kernels, Trilinos, etc.
- Low precision preconditioners can be used to accelerate iterative solvers
  - Preconditioners need to adapt their precision to numerical requirements
  - The precision of the preconditioner determines how much accuracy is preserved

Multi-precision: take-aways (2/2)

(from Anzt and Luszczek webinar)

- For memory-bound preconditioners, decoupling the arithmetic precision from the memory precision provides the runtime savings while preserving a constant preconditioner
- To increase the performance benefits, shift most of the work to the low precision preconditioner
- Mixed precision iterative refinement is a powerful strategy to accelerate linear solves
  - Iterative inner solver e.g. for sparse systems
  - Direct inner solver e.g. for dense systems
- Mixed precision iterative refinement can also be used for eigenvalue problems
  - Low precision eigenvector approximations as input
  - Convergence in high precision after 3-4 IR steps
- The performance benefits depend on the problem and hardware capabilities

Thank you !