Objective Driven Computational Experiment Design: An ExaLearn Perspective

Francis J. Alexander[1] and Shantenu Jha[1,2]
[1] **Brookhaven National Laboratory**
[2]**Rutgers University**

Overview

A fundamental problem that currently pervades diverse areas of science and engineering is the need to design expensive computational campaigns (experiments) that are robust in the presence of substantial uncertainty. A particular interest lies in effectively achieving specific objectives for systems that cannot be completely identified. For example, there may be "big data" but the data size may still pale in comparison with the complexity of the system, or the available data may be scarce due to the prohibitive cost of the relevant experiments.

In current practice, the methodologies by which experiments inform theory, and theory guides experiments, remain ad hoc, particularly when the physical systems under study are multiscale, large-scale, and complex. Off-the-shelf machine learning methods are not the answer—these methods have been successful in problems for which massive amounts of data are available and for which a predictive capability does not rely upon the constraints of physical laws. The need to address this fundamental problem has become urgent, as computational campaigns at pre-exascale, and soon exascale, will entail models that span wider ranges of scales, represent richer interacting physics, and inform decisions of greater societal consequence.

To facilitate the design of computational campaigns across multiple scientific domains, diverse objectives and measures of robustness, we are developing a computational capability for **objective driven experimental design** (ODED) using RADICAL-Cybertools as part of the recently funded DOE ECP Co-Design Center "ExaLearn". This ODED framework will support the integration of scientific prior knowledge on the system with data generated via simulations, quantify the uncertainty relative to the objective, and design optimal experiments that can reduce the uncertainty and thereby directly contribute to the attainment of the objective.

Importance of ODED at Exascale

Although ODED has been always been important, its significance increases drastrically at the exascale. First, computational resources are too expensive for the design of computational campaigns to be conducted in an ad hoc fashion and for the resulting data not to be exploited to its very fullest. Second, is the need to enhance if not preserve computational efficiency at exascale. The ability to apply high-performance computing capabilities at scale, leads to the possibility of greater inefficiency in computational exploration. For example, greater computational capacity might generate relatively greater correlations, and thus less independent data, or lesser sampling.

The interaction between models and data occurs in two directions: (i) The problem of how to use multi-modal data to inform complex models in the presence of uncertainty, and (ii) How, where, when, and from which source to acquire simulation data to optimally inform models with respect to a particular goal or goals is fundamentally an optimal experimental design problem. Creating the conceptual and technological framework in which models optimally learn from data and data acquisition is optimally guided by models—presents significant challenges systems of interest are complex, multiscale, strongly interacting/correlated, *and* uncertain. These challenges must be overcome to realize the promise of efficiency and effective exascale computing. Our framework is designed to ideally support the mathematical developments while being agnostic of any specific formulation.

Example Science Driver: Objective Driven Computational Drug Design

The strength of drug binding is determined by a thermodynamic property known as the binding free energy. One promising technology for estimating binding free energies and the influence of protein and ligand composition upon them is molecular dynamics (MD). A diversity of methodologies have been developed to calculate binding affinities; MD sampling and blind tests show that many have considerable predictive potential. With the demands of clinical decision support and drug design applications in mind, several computational protocols to compute binding free energies have been designed. Different protocols (algorithmic methods) typically involve compounds with a wide range of chemical properties which can impact not only the time to convergence, but the type of sampling required to gain accurate results [1,2]. The advantages of determining optimal computational campaigns include: (i) Greater sampling and higher throughput of drug candidates; (ii) more accurate binding affinity calculations, and (iii) Efficient resource utilization.

Mathematical Formulation of Objective Driven Drug Design Campaign: We believe the problem of determining an optimal computational campaign for a given objective (O) under a given constraint (C) can be formulated as: Imagine there are M different stochastic algorithms which calculate the same quantity of interest but with different variances (errors) for the same amount of computational resource. Conversely, the distribution of run times (t) for a given algorithm at a prescribed variance level (σ) on a given computing resource R is given by $P_R(\sigma,t)$. A priori we dont know this distribution $P_R(\sigma,t)$, but we can learn it from ongoing experiments. One possible constraint (C) could be a fixed amount of overall computational resources available to run the algorithms (which can be run in parallel); another constraint could be to get the "optimal" answer in a given wall clock time. An objective could be to find the selection of algorithms so as to minimize the variance of the estimate of the mean, for the given constraints.

Software System for High-Performance Objective Driven Experimental Design

To promote interoperability and reuse across different scientific problems, objectives and optimization criterion, the software systems for high-performance optimal experimental design must be architected and implemented to support diverse usage modes, different combinations of capabilities with minimal customization, refactoring or new development.

At Brookhaven National Laboratory, we are developing the ODED framework to determine "optimal" surrogates for expensive cosmological simulations, to optimize computational campaign configurations for drug discovery and optimal model selection for materials and climate science probelms. The ODED framework is being developed in collaboration with RADICAL Laboratory at Rutgers and it leverages the RADICAL-Cybertools – a Building Blocks (BB) approach for HPC middleware [3]. It will utilize existing BB for sophisticated and scalable management and execution of ensemble and optimization style workflows.

References

[1] J Dakka, K Farkas-Pall, V Balasubramanian, M Turilli, S Wan, D Wright, S Zasada, P Coveney, S Jha: Enabling Trade-offs Between Accuracy and Computational Cost: Adaptive Algorithms to Reduce Time to Clinical Insight. CCGrid 2018: 572-577

[2] Concurrent and Adaptive Extreme Scale Binding Free Energy Calculations. J Dakka, K Farkas-Pall, M Turilli, S Wan, D Wright, P Coveney, S Jha: published in IEEE eScience 2018 (arXiv 1801.01174)

[3] M Turilli, A Merzky, V Balasubramanian, Shantenu Jha: Building Blocks for Workflow System Middleware. CCGrid 2018: 348-349