AI RFI Responses, October 26, 2018

Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan RFI Responses

DISCLAIMER: The RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government, National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence (AI), NSTC Subcommittee on Machine Learning and AI, NSTC Subcommittee on Networking and Information Technology Research and Development (NITRD), NITRD National Coordination Office, and/or any other Federal agencies and/or government entities.
We bear no responsibility for the accuracy, legality or content of all external links included in this document.

RFI OSTP-NSTC
California Institute of Technology

Maria Spiropulu, Shang-Yi Ch'en Professor of Physics

October 26, 2018

1 Executive Summary

The High Energy and Particle Physics Experimental program at Caltech is focused on answering fundamental questions about the composition of the universe at the level of the most elemental particles of matter, and how they can help us understand the intricacies of space and time. In doing so, we also probe the farthest reaches of the universe seeking out the nature of dark matter and any connections with dark energy. The program includes pioneering research on the detailed characterization of the Higgs boson and searches for physics beyond the Standard Model (e.g., supersymmetry, dark matter, new force carriers) at the highest energy and intensity hadron collider (LHC and HL-LHC), on neutrino science, specifically the elucidation of lepton masses and flavor mixing through precision measurements of neutrino properties, on searches for new physics at very high mass scales through high-precision measurements (e.g, neutrinoless conversion of muons to electrons); searches for violation of charge-parity symmetry in leptons, with implications on the evolution of the early universe. The program includes a strong instrumentation component with development of novel detector, electronics, and computation technologies. The group has strong collaborative ties and partnerships with the Jet Propulsion Laboratory as well as other National and International Laboratories, including Fermilab in the US and CERN in Europe. The group is also pioneering research and applications in emergent intersecting areas with nuclear physics, materials science and relevant directions of quantum information science. The Caltech HEP group is also involved in high-performance computing (HPC) and high-throughput software-defined networking (SDN) as well as distributed or grid computing. In all the areas of our fundamental physics research we find critical utility and application for AI methods, as well as a dire need for **further fundamental and applied research into AI, specifically in the areas of fundamental machine learning problems like interpretability, bias, controllability, visualization and causality.** The leading industry efforts on AI are under-investing (at best) in these topics, and there is a pressing need for technical progress and leadership. We believe that through AI research for science, such challenges can be addressed in a transparent and significant way.

As we seek to use AI to derive new knowledge and understanding about the physical world, it is crucial for the AI methods to be interpretable by humans and able to build on prior physical knowledge. Therefore, we see the

development of models that are able to reason about the physical world in a causal way as high-priority, high-impact direction of research. The HEP program is particularly well-placed to participate in research in causal modelling, as we have at our disposal vast amounts of data backed by well-established quantitative models based on experimentally verified physics, namely the standard model (SM) of particle physics.

It is important that provisions are made to publish the full datasets used for any new AI developments in our field, such that the growing AI research community can be engaged. We envision a two-pronged approach of continuing programs such as the Compact Muon Solenoid Open Data program and related initiatives to release the full datasets collected and simulated by the experiments, as well as releasing specific benchmark datasets and models for specific problems of high interest such as particle tracking, b-quark identification or Higgs boson decay identification to enable cross-validation of new results by a wider community of researchers.

We find strategies 1 (fundamental research AI research into causal models), 5 (shared datasets) and 6 (standards and benchmarks) to be among the most important directions for future AI research and believe investment in these will have wide-ranging impact on using AI for knowledge discovery and beyond.

2 Introduction

In this reply, we provide the viewpoint and comments of the Caltech Compact Muon Solenoid HEP group to the Request for Information on the 2016 AI R&D Strategic Plan.

The Caltech group is actively pursuing data-intensive fundamental research on high-energy physics with AI methods and deep learning in particular used throughout. Fundamental research into AI for knowledge discovery at the Caltech group is made possible by the unprecedented scale and complexity of the data generated at the LHC and recorded by the multipurpose detector experiments, coupled with the uniquely detailed and predictive underlying physics models arising from quantum field theory accessible through well-tested theory and simulation. This combination is specific and unique for data-intensive physical sciences and can propel the development of AI methods that go beyond a statistical description of the results, with the goal of incorporating causality and learning effective models from fewer data points.

3 Summary of AI Group Activities

Using the latest data from the CMS experiment at the LHC, we study the nature of the Higgs boson discovered in 2012, look for signatures arising from possible BSM physics such as SUSY or lepton flavour violation, develop reconstruction and analysis methods for anomaly identification to look for processes which may otherwise escape detection, develop new methods for the identification of signatures from heavy flavoured quarks and the rejection of pileup arising from background processes and develop distributed computational methods for the analysis workflows. Below we describe research activities in the group that employ AI methodologies. For the program described below we are collaborating and interacting at various levels with machine intelligence groups at Fermilab, CERN, DeepMind, JPL and Caltech Computer Science colleagues, as well as various hardware industry vendors (such as NVIDIA, IBM, Unity3D and Microsoft to mention a few).

Higgs analyses Detailed studies of the Higgs boson discovered at the LHC are among the highest priority deliverables for Run 2 of the LHC, as ultimately knowledge about the interactions of the Higgs boson and the shape of the Higgs potential will shed light on the Universe evolution and the energy scale of BSM physics. Very recently, with major contributions from our group, the Higgs has been observed to interact with top quarks. The relevant analyses involve combinations of classification DNNs based on signal and background process simulation, as well as theory-driven statistical analysis based on physics models. The HEP community has reacted to this work and developed parallels between the ML and theory-driven approaches that may result in more powerful ML techniques for the study of processes where simulation samples can be complemented with theory. We continue developing these methods for the upcoming di-Higgs flagship analyses that is expected to probe the shape of the Higgs potential, as well as the Higgs decay to muons, which will confirm the mass generation mechanism for fermions. In an ML approach for these analyses the precise understanding of systematic bias introduced by the ML selection is crucial in order to extract physically meaningful results. This project depends on progress in fundamental AI research, open datasets and cross-validation.

Anomaly detection In recent years, the variational auto-encoder (VAE) has been widely used as a powerful deep generative model in various challenging tasks, such as anomaly detection, video generation, and image compression, due to its ability to capture rich probability distributions from data.

The model, consisting of an encoder and a decoder, aims to maximize the marginal log-likelihood of given data with respect to the model's parameters. Many extensions to this model have been proposed to enrich the distributions in the latent space. In our group, we extend this method to LHC-scale datasets.
The goal of this project is to search for new physics at the LHC by training a VAE on known physics processes and developing a statistical test for isolating outlier events not predicted by common BSM models for further analysis. Besides advancing the LHC discovery program, we have the potential to benefit AI research more widely by developing autoencoder methods on complex data with a causal structure not found in many other benchmark datasets.

Pileup simulation, mitigation In order to maximize the collected luminosity and thus the effectiveness of the LHC, the machine is operated in a mode where on average 40-80 simultaneous and independent proton-proton interactions take place during one bunch crossing of the collider, with the number of interactions (pileup) expected to increase to 200 in the HL-LHC phase of the project. Most of these proton-proton collisions happen at low transverse momentum and represent well-understood interactions from QCD, therefore they must be effectively removed in the data analysis step. Classical rule-based removal methods perform adequately at low pileup occupancy, but scale poorly for the future operating conditions. Recently, we have demonstrated the usefulness of ML methods based on a novel graph representation of the collision data, which allow the interesting high transverse momentum interactions to be filtered with significantly higher efficiency compared to traditional methods. Graph representations of the data and graph neural networks (GNN) allow features of the underlying data-generating physics process to be taken into account in the ML method. Similarly to the advances of computer vision by using convolutional neural networks (CNN), ML methods that are able to encode symmetries of the underlying system, such as GNNs, can have wide-ranging effects on the physical sciences.

Double-B-tagger In searches for new physics at the LHC it is critical to distinguish single-jet objects that originate from the merging of the decay products of Higgs bosons decaying to heavy-flavour quarks at high transverse momenta, from jets initiated by single partons that are characteristic of background processes. The main direction of research in this area is to use DNNs to improve the efficiency and better suppress the backgrounds in signatures

involving Higgs boson decays to b-quark jets at the CMS detector. This involves developing new neural network architectures for the classification problem, taking into account the fundamental physics of the data-generating process. We develop further an interaction network (IN) architecture (originally proposed by Google DeepMindfor learning about simple physical objects and relations) as an experimental neural network tagger. The use of the IN at the LHC further allows the complex system of interacting particles to be decomposed in terms of individual objects and the relations or interactions between the objects. The IN is a generalization of a graph-based neural network with integrated physics simulation. Our research into implementing an interaction network provides the first working model of a general purpose, learnable physics engine in particle physics. It will be crucial to validate this method more widely on publicly-available datasets.

HPC optimization The Caltech group operates an LHC Tier-2 computing center with around 300 high-performance production-grade servers consisting of around 8000 CPU cores, 15 TB of memory, 5 PB of hard disk space made available to the US LHC community and the LHC researchers more widely. All the components of the data center generate a flow of logging and metrics data which is recorded in a time series database and can be correlated to overall performance of the system as measured by well-established standardized testing, as well as measurements of the site performance in terms of number of LHC events processed and efficiency of resource usage. In addition to optimizing the energy efficiency and thus cost effectiveness of the data center,this rich data flow allows the work in the datacenter to be optimized in various ways, either by predicting or mitigating imminent hardware or software failures or by scheduling workflows between servers and between users to better optimize the utilization of the cluster.

Smart Switches Multipoint computer networks such as the worldwide LHC Grid or the Tier-2 internal network experience resource contention when several accessors seek to use a limited amount of bandwidth by oversaturating network links. Traditionally, custom heuristics have been used to optimize specific networks, however, as any model-based approach, they tend to be static, overly specific and dependant on highly accurate input data. By taking advantage of modern network switching hardware such as the Barefoot Tofino (smart switch developed by AT&T), large amounts of flow metrics can be collected and acted upon in real time. In our group, and in collaboration

with the Palo Alto Foundry and Fermilab, we are investigating the possibility of using deep learning to optimize network flows in conjunction with such smart switches, using the Tier2 and local networking expertise as a test-bed. This involves the prediction of future network flows based on past data in a time series of flow rates across network ports, as well as a control theory and reconfiguring the network dynamically for optimal resource usage. This may lead to the development of future networking hardware able to dynamically adapt based on past flows without manual intervention or reconfiguration.

AI Computing Operation Due to large amount of data produced at the LHC and the subsequently large dataset of simulated events required to perform analysis of the data, there is the need for a large set of computing resource in the form of the LHC grid. The LHC-grid will grow significantly in the next decade, primarily by addition of HPC resources to the network of computing centers contributing. Any distributed computing system is bound to temporary failure that can lead to significant amount of failed workload which currently is reviewed and operated on by a group of human operators. Our work within the CMS computing infrastructure aims at automatizing this task to the highest possible level, with the use of AI to learn from the operator and produce useful prediction of actions to be taken. This work is carried out within the field of Human/AI interaction and is crucial to realizing the LHC physics program at the horizon of the HL-LHC and further more to provide efficient usage of HPC resource used within the context of HEP.

Deep Learning Distributed Training and Optimization Training deep neural networks using the stochastic gradient descent technique is a computing intensive task rendered tractable by using general purpose graphic units (GP-GPU). Despite significant speed up, training a full network to convergence can raise to several days. We have developed an mpi based distributed training framework for the two major deep learning pythong framework (Keras and pytorch) to harvest the full capacity of our computing servers with up to 8 GPU per node. We have published and bench-marked this software at various supercomputer facilities displaying a fair scaling with the number of nodes in use. We have extend our framework to perform hyper-parameter optimization of neural networks. This optimization can use up to several thousands of nodes and take advantage of the exa-scale computing facilities in the US such as ORNL and others.

Optimizing Simulation with Evolutionary Algorithms Existing tunes for Monte Carlo event generators are commonly the result of a time-consuming manual effort that requires extensive expert knowledge. Since the discovery of new physics relies on discrepancies between simulated and experimental data, it is crucial to tune event generator parameters to minimize error against existing data, thus increasing the visibility of new phenomena. We have developed an automated framework in which existing Monte Carlo tunes are improved by evolutionary algorithms, enabling fast tuning on arbitrary experimental datasets and observables. We compare the performance and evaluation time of several evolutionary algorithms and benchmark them against Bayesian optimization. Ultimately, our results show that the combination of evolutionary algorithms with existing tunes can speed up the tuning process of future event generators, as well as providing a general yet effective method for creating detector-specific tunes. We plan on going further with this technique with data from the LHC for tuning of CMS simulation.

Charged Particle Tracking with Deep Learning This is work based on the HEP.TrkX DOE/ASCR/HEP project, assembled as a consortium of LBNL, Fermilab and Caltech. The project aims at exploring applications of machine learning and deep learning to the challenge of charged particle tracking in the horizon of the High Luminosity LHC (HL-HLC). The currently used algorithm for charged particle tracking are using Kalman filtering techniques to perform the pattern recognition of particle trajectories through clouds of hits deposited in the detector. These algorithms are scaling worse than quadratically with the hit density, and are known to reach a computation limitation for the HL-LHC. We have developed several models to perform tasks of charged particle tracking and we are continuing the investigation using the scaling of graph neural network approaches. Due to the hit density in the detector expected for the HL-LHC era, the models we are training are growing in size and require a large amount of computing resource to converge. We aim at leveraging our knowledge of distributed training at exa-scale super computing facilities in the US to arrive to our goals.

Quantum AI While the topic of Quantum AI is covered under the National Quantum Initiative, benchmarking of Quantum AI developments rests within AI efforts. In our group, we have demonstrated that we can use existing quantum annealers to speed up the optimization of classical machine learning classifiers. This was achieved by recasting the learning process of a Higgs physics problem, as a quantum annealing problem of finding the

ground state of an Ising spin model, solved using a D-Wave quantum computer of a thousand qubits. Although we provided only a proof of concept so far (published in Nature), using quantum annealing would possibly allow the ML classifiers to be optimized using less data and with more resilience to overtraining compared to classical approaches, while still being interpretable. Additionally, quantum annealing that has been demonstrated to outperform classical simulated annealing may be useful for other optimization problems, including tracking at the HL-LHC and computational biology.

4 Recommendations for the Strategic Plan

We strongly support the high-risk, high-reward long-term investment in fundamental AI research (Strategy 1). It has been pointed out that moving towards general or strong AI that is able to learn solutions to a large variety of tasks from a comparably small amount of examples (as is the case for biological intelligence) may well require causal modelling of the world beyond a statistical analysis of correlations and associations. As a specific example, it has recently been demonstrated that introducing physics-informed constraints in the optimization process can result in faster and more data-efficient learning. We find that scientific data such the LHC data, can provide benchmarking for such developments.

As we develop solutions to the computational challenges of the LHC from AI, for example in addressing the problem of scaling multidimensional particle trajectories tracking to high-occupancy environments, it is vital to analyze the computational costs from the perspective of algorithmic complexity and further improve our understanding of the capabilities and limitations of AI solutions, taking into account available and upcoming hardware such as highly parallel GPU or FPGA devices or biologically-inspired platforms such as neuromorphic chips. We work together with industry partners such as nVidia to benchmark recent GPU devices in our computing facility.

The LHC project, serves as a testbed for the development AI methods where very large amounts of data have to be reduced to meaningful conclusions and measured physical quantities with systematic biases that are minimal and understood. Interpretability, fairness and a reduction of bias are therefore necessary for any AI-based approach to provide meaningful knowledge about nature. We plan to collaborate with researchers from the AI domain to guarantee this for ML applications at the LHC. Concretely, we

will establish benchmark datasets for the physics use-cases so any proposed AI methods can be cross-validated and improved independently by the scientific and AI community. In comparison with many other fields, the data collected at the LHC should not be subject to significant privacy or copyright constraints, thereby making it possible to share with a wide community, as demonstrated by the CMS Open Data project that has been engaging researchers from outside the physics collaboration for original research in AI. As already done through the Kaggle HiggsML challenge and the TrackML challenge we plan to further bring in expertise from the AI research community for the domain-specific data of the LHC and therefore foster useful collaboration through open datasets with relevant metadata and descriptions. Therefore, we believe that furthering Strategies 5 and 6 is particularly important for the development of AI-based science.

In closing, we find a strong incentive for collaborations across scientific academic and research institutions with government and industry, with target to address pressing challenges in advanced AI research such as interpretability, controllability, fairness, bias and causality. High energy particle physics, among other sciences, offers a valuable testbed due to the computation requirements and massive data it offers. Practically such efforts will call for i) accessibility and availability of the US HPC centers and associated scientific datasets, ii) promoting and funding dedicated exchanges and interactions between data science and domain sciences iii) funding of local R&D computing facilities in academic and research institutions including National Laboratories for prototyping AI techniques and scientific applications iv) targeted workforce development with training students, postdocs and faculty in the sciences to understand, use, benchmark and further develop AI methods produced by dedicated AI industries.