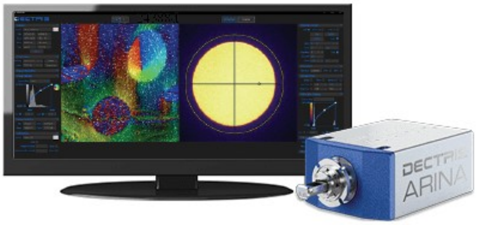
Practical and Parsimonious Real-Time Analysis in Materials Microscopy

Joshua C Agar

DECTRIS

ARINA with NOVENA

Fast 4D STEM



DECTRIS NOVENA and CoM analysis of a magnetic sample.

Sample courtesy: Dr. Christian Liebscher, Max-Planck-Institut für Eisenforschung GmbH.
Experiment courtesy: Dr. Mingjian Wu and Dr. Philipp Pelz, Friedrich-Alexander-Universität, Erlangen-Nürnberg.

Downloaded from https://academic.oup.com/mam/article/29/Supplement_1/1920/7228131 by guest on 13 September 2024

Microscopy and Microanalysis, **29** (Suppl 1), 2023, 1920
https://doi.org/10.1093/micmic/ozad067.992

Microscopy AND Microanalysis

Meeting-report

Practical and Parsimonious Real-Time Analysis in Materials Microscopy

Joshua C. Agar[1]

[1] *Department of Mechanical Engineering and Mechanics, Drexel University, Philadelphia, PA, United States*

*Corresponding author: jca92@Drexel.edu

Downloaded from https://academic.oup.com/mam/article/29/Supplement_1/1920/7228131 by guest on 13 September 2024

Machine learning has tremendous capabilities to advance scientific experiments. There are, however, significant challenges with the practical implementation of these tools. In particular, the absence of machine-interpretable physics-conforming models and computing infrastructure for automated analysis represent significant barriers. Here, we discuss our progress in codesign of scientific experiments, physics-informed machine learning models, and hardware implementations for multimodal and high-velocity microscopy. We will discuss how we address 3 practical problems in data intensive microscopy. First, we will discuss our infrastructure for automated collection and collation of microscopy data. This includes a federated and searchable scientific data management system. This system allows retrospective discoveries through logical queries on schema-free metadata. Furthermore, we show how we can use neural networks to create image similarity projections that allow the creative exploration of large unstructured microscopy databases. Secondly, we will show how physics-conforming neural networks can be developed to improve and accelerate the fitting of materials spectroscopy when governing physics is known. We demonstrate the efficacy of this approach on band-excitation piezoresponse force microscopy and how stochastic averaging in neural networks can significantly improve information extraction from noisy data. We then extend this concept to 4D – scanning transmission electron microscopy strain mapping. We develop a neural network architecture to automate strain mapping analysis. This is achieved by building a cycle-consistent spatial transforming autoencoder where we embed an affine transformation to learn the physics of geometric transformations parsimoniously. We discuss compression, regularization, and optimization methods required to automate this problematic training process. We achieve better performance (0.3 sub-pixel precision) than conventional template matching techniques implemented in py4DSTEM. This highlights the importance of adding physics constraints into neural networks and carefully considering the optimization and regularization methods. Finally, we demonstrate a pathway to deploy these methods for real-time analysis and control. 4D-STEM images can be acquired at frequencies up to 5000 Hz. For control applications, this requires that streaming analysis happens in less than 200 microseconds. This is faster than what can be achieved on general-purpose computing (e.g., CPUs and GPUs). For example, NVIDIA's fastest benchmark from an EfficientNet has a minimum latency of 700 microseconds. Overcoming these challenges requires deploying machine learning models on programable logic (e.g., field-programmable gate arrays [FPGAs]). This, however, is not a straightforward process. FPGAs are resource-limited, requiring that machine learning models are sufficiently compressed to meet hardware and latency requirements. We discuss using model distillation and quantized-aware training with second-order optimizers to create compact machine learning models. We then discuss a tool flow adapted from the triggering system of the Large Hadron Collider where we convert machine learning models trained in QKeras (the quantized-aware Keras package) to high-level synthesis (HLS) using HLS4ML. Once in HLS, we compile the model for Xilinix hardware targets. We demonstrate simulated streaming inference (including I/O) at <100 microseconds per inference, sufficiently fast for real-time edge strain mapping. Ultimately, this work develops methodologies to seamlessly integrate AI systems with voluminous, high-velocity, and noisy experimental data.