AI/ML/Autonomous Experimentation for Thin Films Processing

Room 317 - Session AIML3-FrM

Connecting Datasets, Models, and Characterization

Moderators: Ms. Molly McDonough, Pennsylvania State University, **Dr. Amanda Sanchez**, Universidad Anahuac

9:45am **AIML3-FrM-7 In the Design of Thin Charge-Transport Films for Perovskite Solar Cells, Predictions from Machine Learning and Physics Only Partially Overlap**, *Joseph Oti, Elvis Twumasi,* SUNY Polytechnic Institute*; Nana Derkyi,* SUNY Polytechnic Institute, Ghana*; Tabiri Asumadu, Desmond Klenam, Bernice Abraham,* SUNY Polytechnic Institute*; Rodica Neamtu,* ***Nancy Burnham****,* Worcester Polytechnic Institute*; Iulian Gheraisou, Winston Soboyejo,* SUNY Polytechnic Institute

We present a hybrid machine learning (ML) and physics-based computational modeling framework that informs experimentally realizable thin charge-transport-layers for perovskite solar cells (PSCs). Using supervised ML models, 12294 previously published curated datapoints predicted device efficiency. Ten high-efficiency descriptor combinations with greater than 21% power conversion efficiency were mapped into SCAPS-1D (a physics-based simulation software) employing a $FASnI_3$ absorber to assess their physical plausibility. In contrast with the ML predictions, the SCAPS configurations yielded PCEs in the range of 0.03–22.7%. Discrepancies between ML and SCAPS predictions arose from interfacial recombination, optical losses, and physical constraints not explicitly encoded in the ML dataset, describing a tightly focused set of fabrication parameters. Thus, the complementary use of both machine learning and physical simulations offers a time- and material-saving scalable pathway for the development and fabrication of low-cost photovoltaic devices.

10:00am **AIML3-FrM-8 Closed-Loop Ai Workflow for Literature-Guided Perovskite Thin-Film Discovery***, **Jordan Marshall**,* University of Tennessee Knoxville

In this project, we are developing an AI-assisted thin-film materials discovery workflow for halide perovskite solar cells. The goal is to make materials discovery more connected, adaptive, and efficient by bringing together steps that are often handled separately: literature knowledge, predictive modeling, robotic synthesis, and high-throughput characterization. Instead of treating these as independent parts of the research process, the workflow is being designed so that information from one stage can help guide the next. A large part of this work has focused on building a literature-mining agent for perovskite solar-cell papers. The perovskite literature contains a large amount of useful information, including device structures, perovskite compositions, processing conditions, interface layers, power conversion efficiencies, and stability measurements. However, this information is often reported in different ways across papers, which makes it difficult to use directly for machine-learning studies. The literature agent is designed to read papers, extract the relevant information, and organize it into a structured format compatible with the Perovskite Database. The workflow also aims to preserve where the information came from, separate evidence from model-ready features, flag incomplete records, and identify possible duplicate device entries. This makes the dataset update process more transparent and easier to check. The updated dataset can then be used to train sequential machine-learning models for device performance and stability. First, power conversion efficiency is predicted from design, processing, interface, and device-context features while avoiding target leakage from already measured performance values. Stability is then modeled using a similar feature space. We compare design-only stability prediction, stability prediction using model-predicted PCE, and post-fabrication prediction using measured initial device performance. This structure helps separate early design guidance from later experimental decision-making. This work is an important step toward a closed-loop discovery platform for perovskite and related semiconductor thin films. Literature knowledge can update the database, models can suggest promising directions, robotic synthesis can test those ideas, and high-throughput characterization can provide new information that feeds back into the system. While the workflow is still being developed, it provides a practical foundation for more efficient and autonomous thin-film experimentation.

10:30am **AIML3-FrM-10 Multimodal AI Meets Materials Characterization***, **Linda Hung**, Weike Ye, Jithendaraa Subramanian, Daniel Schweigert, Amalya Johnson, Flora Chen,* Toyota Research Institute **INVITED**

Many large-scale materials AI models rely on atomic-resolution crystal structure as a required input. Such models leverage the availability of large computational datasets but can be difficult to apply to experimental workflows. We present multimodal models that are instead built on inputs more commonly available in the lab, such as XRD and composition, and show benchmarks on use cases including noisy XRD and incomplete composition. We then discuss what it takes to build the datasets these models need, and share progress toward that goal.

11:00am **AIML3-FrM-12 Building Informative Materials Datasets Beyond Targeted Objectives***, **Rafael Espinosa Castañeda**,* University of Toronto, Canada, Mexico*; Jason Hattrick-Simpers,* University of Toronto, Canada

Materials science data collection can be expensive, making the reuse and long-term utility ofdatasets critically important for future discovery campaigns. In practice, researchers prioritize a subsetof properties due to research interests. However, ignoring a subset of outcomes in data collectioncampaigns potentially generates datasets poorly suited for future learning tasks. Here, we present a framework for dataset construction that maximizes informativeness for target properties of interest while preserving performance on untargeted ones. Our approach uses diversity-aware selection to ensure broad coverage of the materials space. In noisy experimental dataset construction, we find that without our diversity-aware framework, prediction performance on untargeted properties can degrade by up to sim40% relative to random sampling, whereas applying our framework yields improvements of up to 10\%. For targeted properties, performance can degrade with respect to random sampling by up to 12.5% without diversity, while our framework achieves gains of up to sim25%. Incorporating diversity into dataset construction not only preserves informativenessfor the targeted properties, but also improves materials coverage for potential future objectives. As aresult, the constructed datasets remain broadly informative across considered and unconsidered outcomes,ensuring unbiased quality entries and mitigating cold-start limitations in subsequent modeling anddiscovery campaigns.

Author Index

Bold page numbers indicate presenter