3098
doi:10.1017/S1431927622011540
Microsc. Microanal. 28 (Suppl 1), 2022

Synthetic Data Curation Strategies for Robust Model Development: A Case Study with HRTEM Micrograph Segmentation

Luis Rangel DaCosta*[1,2], Katherine Sytwu[2], Catherine Groschner[1,2], Mary Scott[1,2]

[1] University of California, Berkeley, Department of Materials Science, Berkeley, CA

[2] National Center for Electron Microscopy, Lawrence Berkeley National Laboratory, Berkeley, CA

*luisrd@berkeley.edu

Mounting research evidence has shown that neural network (NN) models can be impressively performant in processing and analyzing data for challenging, niche, and complex scientific tasks, including electron microscopy [1]. As domain scientists begin to more commonly implement machine learning (ML) tools into their scientific workflows, we need to better understand how these relatively black-box stochastic models behave in new environments and how we can make pragmatic and data-informed decisions to ensure that scientific results gleaned from ML tools are valid and trustworthy. While there has been considerable ML research concerning model architectures and optimization, relatively little has been done to understand how data affect model performance, especially in scientific applications. Data curation provides scientists a singular opportunity to use their expert domain training and scientific priors to design ML tools that are well-suited to their needs. Unfortunately, for many tasks, and especially supervised-learning tasks, there is a dearth of suitable and labeled data ready for use in ML model development. For this, we can leverage modern computing resources and generate suitable simulated datasets as a replacement to effectively tackle challenging microscopy analysis tasks [2]. To further address this need in the field of electron microscopy, we have developed a series of tools and programmatic workflows which can be used to automatically generate large simulated databases as well as to statistically assess the performance of NN models under varying data conditions.

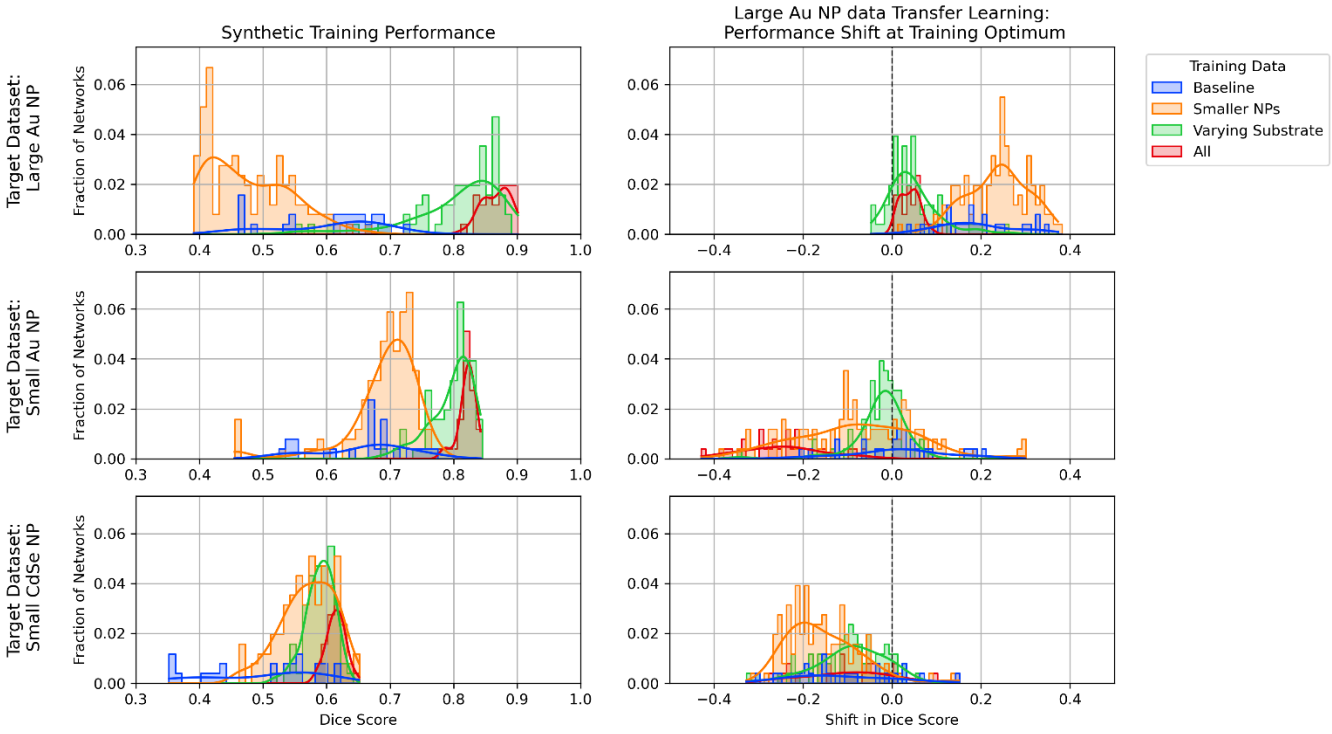
In this work, we present a general framework, some guiding principles, and some automated tools for constructing arbitrary synthetic datasets which closely mimic experimental data as produced on the TEM and for studying the use of synthetic datasets in machine learning workflows. Here, we study a simple model problem, segmentation of nanoparticles in high-resolution TEM micrographs, and analyze the performance of models on several different experimental datasets. We also demonstrate that models trained completely on simulated data can achieve state-of-the-art performance; model performance, both in- and out-of-distribution, can be further saturated with transfer learning on small amounts of experimental data (Fig. 1).

Our data generation pipeline begins with a structural generation tool, Construction Zone (CZ). CZ is an open-source Python package, built on top of popular materials modeling packages [3,4,5], that allows for the generation of arbitrary nanoscale atomic scenes in an algorithmic and automated way. We used CZ to generate several thousands of spherical gold nanoparticles with random planar defects, which are then placed at random orientations and random locations onto amorphous carbon substrates. For each structure, with and without the substrate, we perform HRTEM simulations using Prismatic [6]. Aberrations, arbitrary focal conditions, and Poisson noise are applied to the simulation outputs to generate data under varying imaging conditions. Thresholds of the phase image of the substrate-free nanoparticle are used to create corresponding segmentation masks. Saved data are annotated with

relevant metadata–such as structural details and applied data augmentations–so that arbitrary subsets of the data can be selected and so the data generation process can be fully backtracked; metadata are propagated forward as data are further processed from the structural generation stage.

To understand the effectiveness of data curation strategies on HRTEM segmentation performance, we trained several hundreds of models on various subsets of simulated datasets under fixed training conditions and fixed model architecture. For each unique subset of training data, we trained at least five models to help control for stochastic variance of training performance. Each model was then further trained in a transfer-learning phase on a small portion of experimental data. Performance benchmarks on several experimental datasets were measured at each epoch while training. "Real-time" performance benchmarks on out-of-distribution datasets provide great insight into the generalization dynamics of trained models and are crucial for understanding how NNs can be trained to generalize across applications and the effects of data curation (Fig. 2). Trained models are similarly saved with accompanying rich metadata.

Our results indicate that some aspects of dataset composition might not be so important, such as precise modeling of the structures or aberration conditions, while other aspects, such as structural variation and imaging conditions, can be greatly influential to model performance. Careful data curation can lead to robust model training and performance without needing an extremely large training dataset. Our data pipelines can be trivially extended to much more complex characterization tasks, such as grain identification or combined real-space imaging and spectroscopy, and can easily be scaled to train larger, more generally powerful networks. Metadata rich databases, in this study and future work, are particularly important in facilitating our analysis and enabling us to make (more) precise statistical statements.



https://doi.org/10.1017/S1431927622011540 Published online by Cambridge University Press

Figure 1. Distributions of network performance (left column) on large Au NP data(top), small Au NP data (middle), and small CdSe NP data (bottom) after simulated training on a variety of subsets of simulated data. In general, transfer learning can improve performance (right column); gains in performance are most likely and significant on the target dataset, and still possible on out-of-distribution datasets. Larger dice score is better.

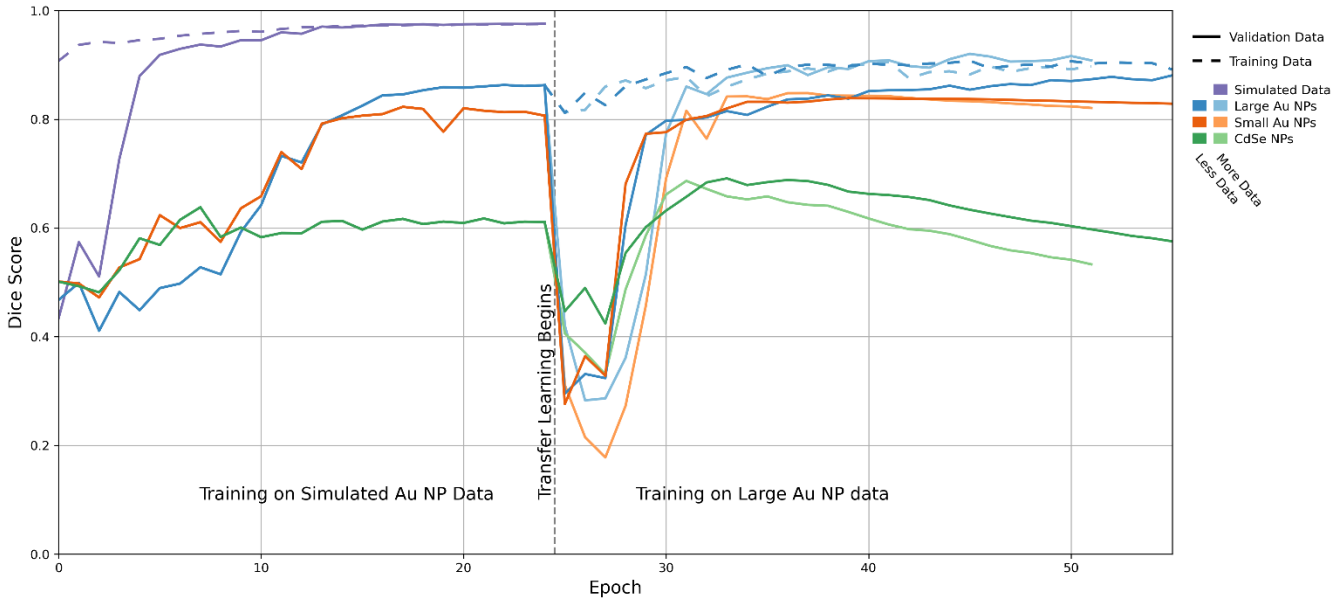


Figure 2. Neural network performance dynamics for network trained on simulated Au nanoparticle data, and then further trained a small amount of experimental data. Performance measured on simulated (purple), large Au NPs (blue), small Au NPs (orange), and small CdSe NPs (green). During synthetic training phase, simulation performance saturates quickly, while experimental performance improves slowly. Performance improves out-of-distribution for a small amount of time during transfer learning, after which it decays—the more data seen during transfer learning (lighter hue), the stronger decay effect. Larger dice score is better.

References:

[1] J. M. Ede, *Mach. Learn.: Sci. Technol.* **2**, (2021), p. 011004. doi: 10.1088/2632-2153/abd614.
[2] J. Munshi *et al.*, *arXiv:2202.00204*, (2022). doi: 10.48550/arXiv.2202.00204
[3] S. P. Ong *et al.*, *Computational Materials Science*, **68** (2013), p. 314–319. doi: 10.1016/j.commatsci.2012.10.028.
[4] A. H. Larsen *et al.*, *J. Phys.: Condens. Matter*, **29** (2017), doi: 10.1088/1361-648X/aa680e.
[5] J. M. Rahm and P. Erhart, *Journal of Open Source Software*, **5** (2020), p. 1944. doi: 10.21105/joss.01944.
[6] L. Rangel DaCosta *et al.*, *Micron*, **151** (2021), p. 103141. doi: 10.1016/j.micron.2021.103141.

https://doi.org/10.1017/S1431927622011540 Published online by Cambridge University Press