

## Benchmark tests of atom-locating CNN models with a consistent dataset

Jingrui Wei<sup>1</sup>, Ben Blaiszik<sup>2</sup>, Dane Morgan<sup>3</sup> and Paul Voyles<sup>3</sup>

<sup>1</sup>Department of Materials Science and Engineering, University of Wisconsin Madison, Madison, Wisconsin, United States, <sup>2</sup>Globus, University of Chicago, United States, <sup>3</sup>University of Wisconsin Madison, United States

Modern electron microscopes can resolve single atoms in images of materials, but experimental images may include significant noise, distortions, and poor contrast. Several convolutional neural networks (CNNs) have been developed to automatically identify atom positions in these datasets. However, each network tends to be developed for a specific problem with unknown applicability on data outside the original test dataset. To overcome these problems, we have developed a benchmark dataset and model performance metrics for testing atom-locating approaches covering a range of image quality and various types of crystal lattices. With this benchmark, we have evaluated the performance of two existing CNN models<sup>1,2</sup>, with both models showing excellent performance over different atomic lattices, except for certain cases like overlapping atomic column images. We find that the key factor that limits a model's performance is the quality of STEM images, as a function of quality metrics including the pixel size, contrast, and image resolution, pointing to the need for community agreement on further optimization of these models and benchmark tests for these and related imaging tools.

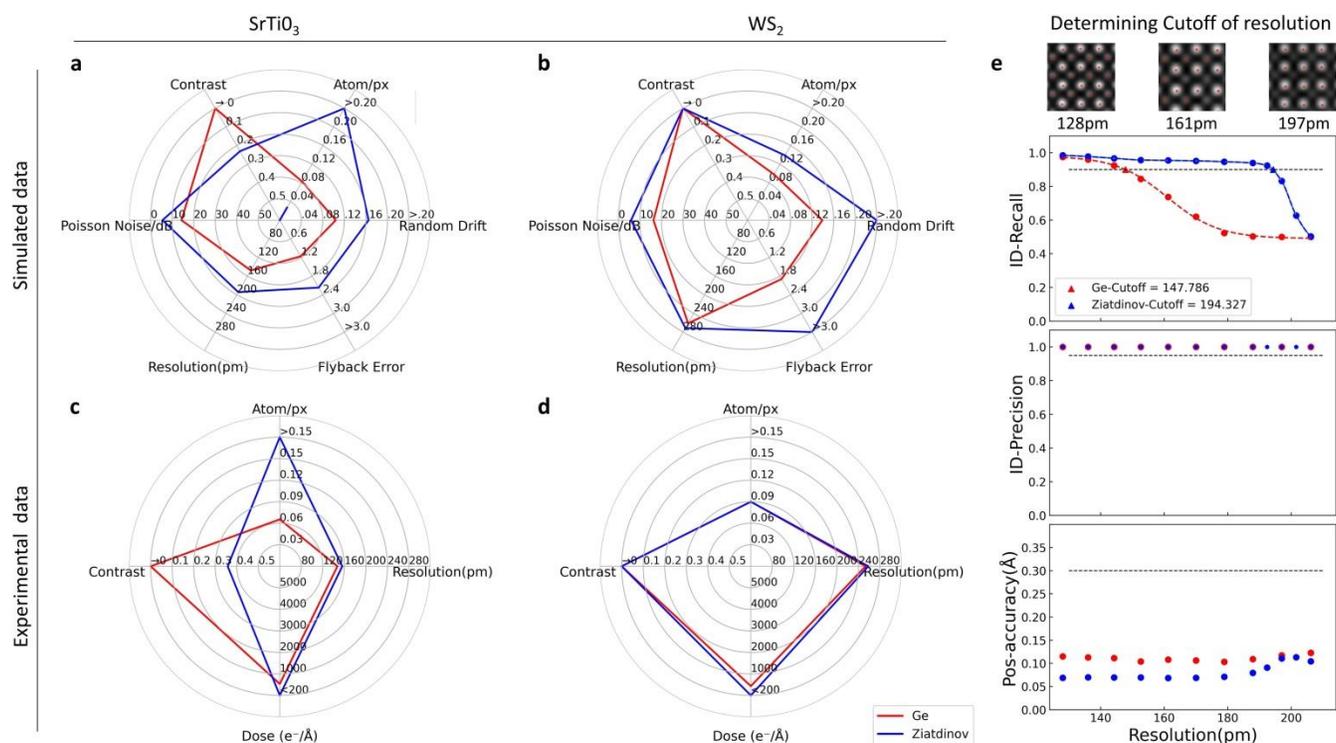
The benchmark dataset contains simulated and experimental HR-STEM images of SrTiO<sub>3</sub> [100] and monolayer WS<sub>2</sub>, which are the same or similar crystal lattices to the training data for the models. Simulated images enable separate, quantitative control over a wide variety of image quality, and experimental images are the ultimate application target. These images are augmented by 34 experimental images covering a wide variety of 2D lattice types, complex unit cells, heterointerfaces, grain boundaries, and other defects. These images are "typical" quality for aberration-corrected HR-STEM. Every image is labeled with the positions of the atomic columns, derived by construction for the simulations and by hand-labeling and refinement for the experiments. We evaluate model performance against an image using ID-recall (TP/(TN+TP)), a measure of the fraction of actual atoms that have been correctly located, ID-precision (TP/(FP+TP)), a measure of the fraction of model predictions that are correct, and position accuracy, the root mean square distance between the correct, model-identified positions and the corresponding ground truth positions.

Here, we have benchmarked two CNN models for atom locating developed by M. Ge<sup>1</sup> and M. Ziatdinov.<sup>2</sup> Both models are based on the encoder-decoder architecture<sup>3</sup> and apply the ability of CNNs to segment and classify objects against complicated backgrounds with high accuracy to locate atoms in HR-STEM images. To quantify each model's performance, we determined a cutoff value for each image quality metric, past which ID-recall < 90%, ID-precision < 95%, or position accuracy > 0.3 Å. Based on observation of multiple test cases within our dataset and general atomic radius, these three criteria fairly satisfy the requirements for extracting accurate information from atomically resolved images. Figure 1(e) shows an example of this process for the microscope resolution metric. The top three images are simulated images as a function of resolution with correctly identified atomic columns marked. The following plots show both models exhibit falling ID-recall because they miss atoms at larger (poorer) resolution, although ID-precision and position accuracy remain excellent.

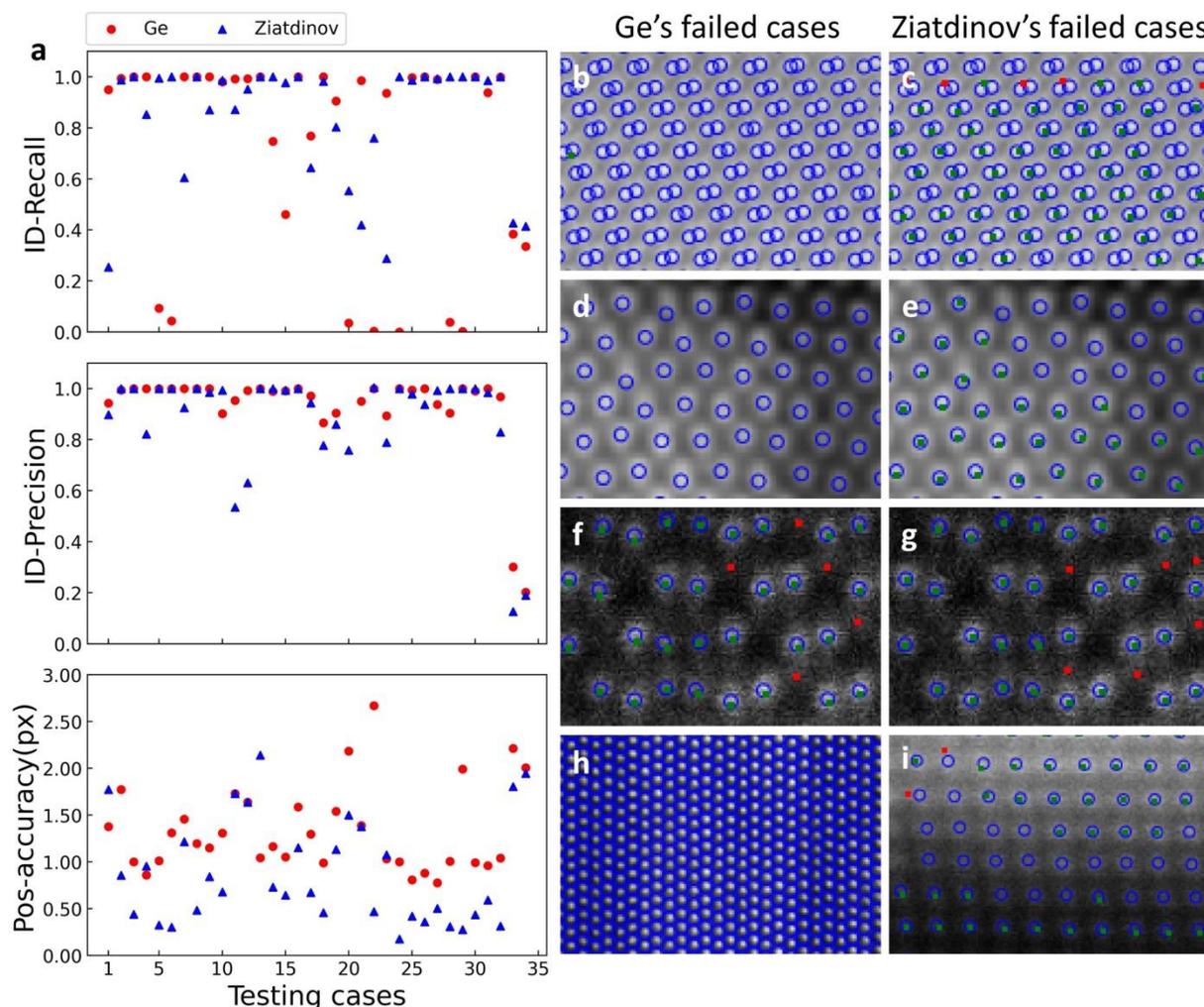
Figure 1(a-d) shows radar plots of the cutoffs for all explored metrics. The larger polygon of Ziatdinov's model means that it has better overall performance than Ge's model on its training dataset 2D-WS<sub>2</sub>. On SrTiO<sub>3</sub> [100] dataset, Ziatdinov's model shows better tolerance to variously degraded images except for contrast. The models generally succeed for high-quality images toward the center of the polygons, except for Ziatdinov's model at very small pixel size, where it incorrectly identifies noise as atoms, decreasing ID-precision, as stressed by the short blue line in Figure 1(a).

Figure 2 shows how the models perform on a few of the more challenging images of other structures. They both fail to find atoms for which the intensity peaks severely overlap (Figure 2b, c). Poor resolution (Figure 2d, e) and poor contrast / high noise (Figure 2f, g) are other common causes of failure. In particular, Ge's model performs poorly on images with large pixel size (Figure 2h), and Ziatdinov's model misses many atoms if the image contains significant contrast variation. We hope this kind of benchmark testing will enable users to select the best model for their images and drive development of even better models.

This research was supported by NSF OAC Awards 1931298 and 1931306



**Figure 1.** Figure 1. Visualization of strengths and limitations of the two models by radar plots over image quality metrics. (a-d): radar plots generated from four groups of test data. (e): examples of images with different resolution, and trends between resolution and three model performance metrics.



**Figure 2.** Figure 2. Generality tests with 34 distinguishing experimental images. (a) evaluation scores of the two models. (b-i) typical failed cases. Blue circles represent ground-truth positions, blue dots are TPs and red dots are FPs of the model.

### References

1. Ge, M., & Xin, H. L. Deep Learning Based Atom Segmentation and Noise and Missing-Wedge Reduction for Electron Tomography. *Microscopy and Microanalysis*, 24(S1), 504–505 (2018).
2. Ziatdinov, M., Dyck, O., Kalinin, S. V., *et al.* Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations. *ACS Nano*, 11(12), 12742–12752(2017).
3. Badrinarayanan, V., Kendall, A., & Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481–2495 (2017).