ARTICLE OPEN

Ensemble learning-iterative training machine learning for uncertainty quantification and automated experiment in atom-resolved microscopy

Ayana Ghosh^{1,2}, Bobby G. Sumpter¹, Ondrej Dyck¹, Sergei V. Kalinin¹✉ and Maxim Ziatdinov^{1,2}✉

Deep learning has emerged as a technique of choice for rapid feature extraction across imaging disciplines, allowing rapid conversion of the data streams to spatial or spatiotemporal arrays of features of interest. However, applications of deep learning in experimental domains are often limited by the out-of-distribution drift between the experiments, where the network trained for one set of imaging conditions becomes sub-optimal for different ones. This limitation is particularly stringent in the quest to have an automated experiment setting, where retraining or transfer learning becomes impractical due to the need for human intervention and associated latencies. Here we explore the reproducibility of deep learning for feature extraction in atom-resolved electron microscopy and introduce workflows based on ensemble learning and iterative training to greatly improve feature detection. This approach allows incorporating uncertainty quantification into the deep learning analysis and also enables rapid automated experimental workflows where retraining of the network to compensate for out-of-distribution drift due to subtle change in imaging conditions is substituted for human operator or programmatic selection of networks from the ensemble. This methodology can be further applied to machine learning workflows in other imaging areas including optical and chemical imaging.

npj Computational Materials (2021)7:100; <https://doi.org/10.1038/s41524-021-00569-7>

INTRODUCTION

Electron and scanning probe microscopies have emerged as primary techniques for exploring the micro-, nano-, and atomic-scale worlds^{1–3}. Multiple examples of the imaging of materials classes ranging from metals and semiconductors to biological and macromolecular systems are abound, with the microscopic tools becoming the linchpins of academic and industrial laboratories throughout the world^{1,4–6}. This rapid progress in imaging techniques have further transformed many imaging areas from largely qualitative to quantitative. Traditionally, in atomically resolved scanning transmission electron microscopy (STEM) or scanning tunneling microscopy (STM) the attention of the researcher has been focused on the presence of large-scale morphological features such as surfaces and interfaces, localized or extended defects, with the conclusions on the physics and chemistry of materials driven by these qualitative observations. Comparatively, the progress in high-resolution imaging allowed quantitative information on the materials structure to be obtained, including the positions of the atomic nuclei in STEM, the center of mass of electronic density of states in STM, etc. This information in turn is related to the fundamental physics and chemistry of materials, and several examples of quantitative studies of materials physics from atomically resolved quantitative observations are now available, including mapping of polarization fields^{7–12}, octahedra tilts^{13–16}, and strains in (S) TEM^{17,18} and surface distortions in STM¹⁹.

However, this progress necessitates rapid analysis of the individual images, as well as dynamic data sets obtained during, e.g., temperature^{20,21}, environment^{22–24}, or beam-induced transformations^{25–29} of solids. The goal of such analysis is to transform the data stream from the microscope, ex situ or in situ, into the coordinates and trajectories of specific features³⁰. These features

can be atoms and molecules in the atomically resolved techniques, or larger scale objects such as nanoparticles and nanorods, cells, etc., in mesoscale imaging. Once established, these features can be used for exploring relevant physics, e.g., strain mapping, visualization of order parameter fields such as polarization or octahedra tilts, and subsequently deriving generative physical models. The need for such analysis necessitates two additional aspects. The first is uncertainty quantification, i.e., ascribing ‘a degree of trust’ to the experimental results. The second is the latency (time necessary to make decision) of the analysis, a consideration that becomes particularly important in conjunction with autonomous experimentation^{31–33} and electron beam atomic fabrication^{27,28,34–37}.

Traditionally, this analysis has been implemented using a broad variety of image analysis tools, ranging from simple maxima finding to correlative filters to Hough transform-based techniques, with each domain area evolving a specific set of tools. Most of these methods required extensive tuning of the hyperparameters in the beginning of the analysis cycle, and often require operator input throughout the process. As such, analysis of more than several images has been unusual. The breakthrough in this area, as in many other areas of computer vision, is the broad implementation and availability of deep learning (DL) networks, pioneered by AlexNet³⁸ and evolving to the large families of ResNets³⁹, U-Nets⁴⁰, etc. In STM, the application of DL for image analysis was pioneered by Ziatdinov et al. for molecular resolved imaging⁴¹ and Wolkow for atomically resolved imaging⁴², and in STEM by Ziatdinov et al.⁴³. This initial effort has grown exponentially in recent years^{44–46}.

However, applications of DL for atomically resolved imaging, or, equivalently, the discovery of a large number of similar objects,

¹Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA. ²Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. ✉email: sergei2@ornl.gov; ziatdinovma@ornl.gov

have significant differences from the classical DL applications. Traditionally, DL methods are optimized to perform feature finding on a large number of possible classes, such as 10 digits in the MNIST database or 10 image categories in CIFAR databases, with multiple examples in each class. Similarly, these are associated with large variability within the class. Comparatively, atomic and particle finding problems typically necessitate finding (almost) identical objects, whereas the imaging conditions can vary between dissimilar experiments or between simulations and experiments. Consequently, the applications of DL in experimental data analysis have to deal with significant out-of-distribution effects. Retraining of the network for a different parameter set is time consuming, both on the labeling and training sides. This aspect is particularly limiting for an automated experiment where the analysis must rapidly adapt to changes in imaging conditions. At the same time, a model trained to account for a broad distribution of experimental parameters may fail to recognize multiple subtle atomic features within a single experimental dataset. In addition, the scientific applications of DL require meaningful uncertainty estimates, which has been a challenge for applications of classical DL to real-world data⁴⁷.

Here, we introduce an iterative DL workflow that surpasses these limitations. This approach utilizes the heavily degenerate nature of high-resolution microscopy, in which a relatively small number of feature classes are possible, and even within a single image multiple realizations of the class are present. We also utilize the fact that in a deep learning a model's final state can be very sensitive to random initialization of weights and shuffling of training mini-batches. These two effects are harnessed in a workflow combining ensemble learning (EL) by multiple neural networks, allowing for selection of artifact-free features and pixel-wise uncertainty maps, and iterative training (IT) where the discovered features are used to retrain the network, focusing its attention on features present in the (heavily degenerate) data and thus increasing the detection limit of the network on the dataset (s) of interest. This allows revealing previously unrecognized features, and to compensate for out-of-distribution drift during the experiments.

RESULTS AND DISCUSSION

ELIT workflow

The classical DL workflow consists of preparing a single labeled training set, selecting appropriate neural network architecture, splitting the prepared training set into 2 (train + test) or 3 (train + test + 'holdout') parts, and tuning the training parameters until the optimal performance on the test and/or 'holdout' set is achieved. Once trained, the neural network is expected to generalize to new, previously unseen data. In the experimental sciences, where the labeling process typically requires an in-depth domain expertise and as a result, the availability of labeled data is usually very limited, one can sometimes use ab initio simulations to prepare a training set. The common challenge for real-world applications of DL (that is, application to a stream of new unfiltered data beyond the test and 'holdout' sets) are the out-of-distribution effects such as changes in data acquisition and processing (inside of an instrument) parameters. An example is the failure of state-of-the-art DL models trained to detect pneumonia in chest X-ray data to provide accurate results on new hospital's data due to the small variations of instrumental and data acquisition parameters⁴⁸.

To set the context we consider the application of DL to experimental imaging of atoms. For DL-based image analysis, the labeling of training data can be performed on the level of individual images (image0001 is 'structure A', image0002 is 'structure B', ..., image6999 is 'structure B') or individual pixels (a pixel at coordinate (i_1, j_1) belongs to 'structure A', a pixel at

coordinate (i_2, j_2) belongs to 'structure B', etc.). The latter is referred to as a DL-based semantic segmentation and will be the focus of our work. Here we explore a very common scenario when there is no labeled experimental data, and one must train a DL model using simulated or synthetic data. The trained model is then used to obtain predictions for the experimental data that typically contain structures and distortions not (fully) considered in the simulations. A similar approach can be applied to situations where the DL model(s) are trained on data from one experiment and applied to new previously unseen experimental images obtained under different conditions.

Our goal is to categorize every pixel in an image as belonging to an atom (or a particular type of atom) or to a background. In this case, the semantic segmentation of experimental atom-resolved images removes all the noise and returns a set of well-defined blobs (corresponding to atoms) on a uniform background where the centers of the segmented blobs correspond to atomic positions. The overall approach is summarized in Algorithm 1 represented by Fig. 1 and schematically depicted in Fig. 2a. It starts with using simulated data to train an ensemble of models instead of just a single model. Recent works have demonstrated that classical and Bayesian EL can lead to improved robustness of DL models and provide meaningful uncertainty quantification under dataset shift and (potentially) for out-of-distribution data^{49,50}. Here, we combined classical EL⁵¹, where each new model is trained with different random initialization of weights and different random shuffling of training data, with a stochastic weight averaging (SWA) procedure⁵², which averages multiple points along the trajectory of stochastic gradient descent at the end of training. This approach is similar to multi-SWA and multi-SWA-Gaussian approaches described by Wilson et al.⁵⁰, but uses a constant learning rate and early stopping instead of a cyclic learning rate. Generally, we found that for simulation-to-real-world-transitions (i.e., model trained on simulated data is applied to experimental data), the role of EL is to identify an artifact-free model or a subset of models, that is, a model or models that do not 'find' unphysical features in experimental data or at least a portion of experimental data. For experiment-to-experiment transitions (i.e., model trained on a subset of experimental data is applied to the remaining data), the EL can be used for improved robustness and for providing uncertainty estimates of predicted values for each point (pixel).

Algorithm 1 Steps to train in ELIT framework

1: $E_{\text{initial}} \leftarrow \text{EnsTrain}(D_{\text{sim}})$	▷ Train ensemble of models
2: $P_{\text{Ens}} \leftarrow \text{EnsPredict}(E_{\text{initial}}, D_{\text{exp}})$	▷ Apply ensemble to expt data
3: $E_0 \leftarrow \text{ArtifactFree}(E_{\text{initial}}, P_{\text{Ens}})$	▷ Choose artifact-free models
4: for i in $\{1, \dots, N\}$ do	
5: $D_i \leftarrow \text{TrainData}(E_{i-1})$	▷ Generate training data
6: $E_i \leftarrow \text{EnsTrain}(D_i)$	▷ Train a new ensemble of models
7: done	
8: $P_{\text{Ens}} \leftarrow \text{EnsPredict}(E_{\text{recent}}, D_{\text{exp}})$	▷ Apply ensemble to expt data
Optional Steps:	
9: $D_{\text{MC}} \leftarrow \text{MultiClassData}(E_N)$	▷ Generate multi-class training set
10: $E_{\text{final}} \leftarrow \text{EnsTrain}(D_{\text{MC}})$	▷ Train a new ensemble of models

Fig. 1 Pseudocode for training models the ELIT framework. First (Line 1) an ensemble of models is trained on simulated data followed by applying each trained ensemble model (Line 2) to experimental data. The best (artifact-free) model or a subset of the best (artifact-free) models is then chosen (Line 3) and training data from a 'good' portion of predictions from the best model(s) is generated. The new training data is used for training a new ensemble of models and make predictions on the experimental data using the newly trained ensemble (Lines 4–8), which in turn is used to refine the experimental training set for iterative re-training of the ensemble. As optional steps (Line 9, 10), a multi-class training set can be generated and used to train a new ensemble of models for multi-class classification.

Fig. 2 ELIT framework. High-level schematic of the ELIT approach is shown in (a). **b** In a classically trained network, the training set must be broad enough to allow sufficient variability to detect classes 1 and 2 present in the system. However, a sufficiently broad training set will also potentially misidentify physically impossible class 3, leading to the classical balance between over- and underfitting. In the ELIT workflow, the multiple trained networks are used to identify the features present in the system, and retrain the network using these features. This allows ‘focusing’ the attention of the network on the classes present in the system, thus dramatically improving the recognition.

Each trained ensemble model is then applied to experimental data and the best (artifact-free) model(s) is selected by a human operator (a domain expert). For experimental data with high variability within the xy plane (single image) and/or along the time dimension (movie), there usually tend to be a portion of data for which the selected model(s) shows a sufficiently high detection rate (percentage of identified atoms). This portion of data (with additional pre-processing if necessary) is utilized to generate a new training set, this time from experimental data, that is used to train a new ensemble of models. This process can be repeated multiple times until the detection rate becomes sufficiently high for the entire dataset. Interestingly, we found that the addition of a small amount of artificial noise and distortions to the training data helps increase the detection rate of the retrained models on the remaining portion of experimental data, especially for dynamic data when there are significant structural changes between the first and the last movie frames.

Here, we have chosen two specific cases to showcase the efficacy of the ELIT framework for atom identification: (i) A dynamically evolving 2D graphene system and (ii) Lanthanum strontium manganite with embedded islands of nickel oxide (NiO-LSMO). The graphene movie contains multiple image frames, accounting for the dynamic nature of the structural evolution occurring in the graphene system with point impurities, whereas the NiO-LSMO image is a single image showing the presence of multiple types of atoms and phases.

ELIT on graphene

We start by training an ensemble of 20 models using the multi-SWA method for a U-Net⁴⁰ model architecture. The training data represent 2D images with atoms as two-dimensional Gaussians and corresponding circular masks of the fixed radius. Alternatively, the Multislice method⁵³ can be used for image simulations, although we did not find any significant differences within the context of the DL applications to 2D systems. The images and

masks are generated using coordinates produced by ab initio molecular dynamics (AIMD) simulations performed on a graphene supercell at 300 K (more details in the ‘Materials and methods’ section). During each run of MD simulation, carbon atoms are removed one by one until a few (10) are left in the supercell. Once created, the images and masks are augmented via random cropping, application of Gaussian and Poisson noises with random scales, blurring, varying contrast levels, and zooming-in randomly in the range between 1x and 2x. The augmentation is used to help generalize to real experimental data. Examples of a few snapshots of AIMD simulations and how images/masks are augmented are shown in the Supplementary Materials.

The results of applying several selected models from the trained ensemble to the experimental data are shown in Fig. 3 in a form of semantically segmented maps corresponding to the final layer of individual DL networks. We note that the accompanying interactive Jupyter notebook allows applying all the models in the ensemble to any experimental movie frame. Interestingly, we found a very significant variation in predictions on the experimental data despite that all the models showed nearly the same prediction accuracy ($73.7 \pm 0.3\%$) on the simulated test data (see Supplementary Table 1). This stark contrast in behavior on the theoretical test data and on real-world data is due to the fact that the former comes from the same distribution while the latter is almost always from a different distribution simply because it is impossible to account for all the instrumental parameters when preparing simulated data. One can see that the first (Fig. 3d–f) and second (Fig. 3g–i) models appear to be capable of avoiding the amorphous regions even though the ensemble models were not specifically trained to do so (there was no amorphous phase in the training set). Notice, however, that the second model shows additional ‘foggy’ features in the extended regions of the graphene lattice, which do not have a direct physical interpretation and are considered an artifact. The third model (Fig. 3j–l) does not automatically filter out the amorphous parts but tends to provide more clear results around point impurities. Finally, all the models in the ensemble start producing unphysical features inside the growing hole in graphene (Fig. 3e, f, h, i, k, l), which means that none of the current ensemble models can be used to analyze the entire dataset. We also note that one cannot simply use an averaged ensemble prediction at this stage, as is common in classical DL, since it will average together both ‘physical’ and ‘unphysical’ features in predictions. By the same token, calculating the dispersion in the ensemble predictions, that is, obtaining uncertainty estimates for each point, is meaningless at this stage.

Next, we select the best, or one of the best, ensemble models (as defined by a domain expert) and create a new training set using the artifact-free subset of predictions on the experimental data. Alternatively, one can use an average prediction from a subset of the best ensemble models. For the graphene dataset, the subset of predictions used to create the new training data set is a portion of the movie (first six frames) without a hole. We use the third model from Fig. 3 as our ‘baseline’ model for the IT because it has a higher detection rate for point impurities. The features associated with the amorphous parts in the model output are removed via a patch-based Gaussian mixture model (GMM) analysis, where one first forms a stack of small image patches (here, we used a 48×48 patch size) around each identified ‘atom’ acting as local descriptors and then applies a two-class GMM to the formed stack (see the Supplementary Materials and the accompanying notebook). Due to drastic differences in local neighborhoods of features associated with graphene and those associated with the amorphous part, the GMM unmixing easily separates a class corresponding to the graphene lattice and a class corresponding to the amorphous regions. The features associated with the latter are discarded during the new training set preparation. The new training dataset is augmented by random cropping of experimental image frames and a constructed mask

Fig. 3 Application of ensemble models trained on simulated data to real data. Predictions in a form of semantically segmented maps from different ensemble models for frames 1 (a, d, g, j), 50 (b, e, h, k), and 90 (c, f, i, l) of the experimental movie from 2D graphene with impurities.

into 224×224 patches. In addition, we applied random $\pm 90^\circ$ rotations, horizontal and vertical flips, small scale jittering, and Gaussian noise with a randomly chosen magnitude. No contrast, blurring, or other types of noise/distortions were applied to training data. The new ensemble of models can be trained using the best model from the previous ensemble as a baseline, which can speed up the computation, or entirely from scratch. For the datasets studied in the current work, both strategies produced comparable results.

A new ensemble of model trained on the (augmented) subset of experimental data is then applied to the entire dataset. This time, since both training and test/validation data come from the same distribution, the variation in predictions of different ensemble models is much smaller and we can use a mean ensemble prediction instead of predictions of individual models, which as in the case of classical DL ensembles provides a more accurate prediction. The results for the frames from the middle and end of the movie are shown in Fig. 4a, b. Comparison of these results with the results of the models trained on simulated data (Fig. 3) shows a significant improvement in the quality of predictions after just a single ELIT iteration. We can further add multiple channels to our training masks to allow for DL-based identification of both positions and types of atoms. Here, we used a simple binary threshold for separating intensities extracted at the atomic positions from the mean ensemble prediction but more advanced methods such as mean-shift clustering can also be applied. The resultant masks have two classes: graphene atoms

and point impurities (here, the 'impurity' is an atom with an intensity higher than C). These masks and the corresponding experimental images are used to train another ensemble of models, now using randomly sampled patches from all the frames and applying the same augmentation procedures as in the previous iteration. Here, we changed the final layer of the initial U-net architecture so that it has three outputs instead of only one (and, correspondingly, the sigmoid activation was changed to the softmax one). For the remaining layers, the weights learned in the previous ELIT can be used as a baseline to speed up the convergence (since atoms remain atoms) although we found that retraining from scratch generally leads to a similar result.

The mean predictions of the ensemble of multi-class segmentation models are shown in Fig. 4c, d. The multi-class ensemble can be re-used on similar experimental data, that is, data obtained under similar experimental conditions with roughly the same feature scale (approximate number of pixels per atom) as we show in Supplemental Fig. 4.

Next, we demonstrate that ELIT framework can provide meaningful uncertainty estimates on the level of individual pixels. In Fig. 5, we show the mean ensemble predictions for two sample regions and the corresponding uncertainty for each class computed as a standard deviation of ensemble predictions for each pixel. The larger uncertainty values at the boundaries of predicted blobs are expected as the (fixed) size of 'atoms' in the mask is chosen somewhat arbitrary. At the same time, the presence of 'doubling' (dumbbell-like structures) in uncertainty

Fig. 4 Mean ensemble predictions after re-training with subsets of experimental data. **a, b** Mean DL prediction (semantically segmented map) from the ensemble after the first re-training using only frames from the beginning of the movie. **c, d** Prediction (semantically segmented map) after second retraining using multi-class segmentation masks with training data randomly sampled from the entire movie.

Fig. 5 Estimation of uncertainty in ensemble predictions on the level of individual pixels. **a, e** Experimental sub-images from frames 95 and 90. **b, f** Mean ensemble prediction. **(c, g)** and **(d, h)** show normalized uncertainty maps calculated as a standard deviation of ensemble predictions in each pixel for graphene atoms **(c, g)** and impurity atoms **(d, h)**. See the accompanying notebook for uncertainty maps on the full images. The higher intensity in color in **(c, d)** and **(g, h)** correspond to higher uncertainty associated with the atoms identified in those specific locations.

maps for impurities is interesting and can be related to unstable atoms jumping between different lattice sites during the scan. Note that this information cannot be captured by a single DL model, which would assign an atom to one of the positions without providing the uncertainty maps. We found that about 6–8 SWA models in the ensemble are typically enough for the meaningful pixel-level uncertainty estimation.

ELIT on NiO-LSMO

Finally, we show that the same approach can be used for finding and classifying (nearly) all the atoms in a single image using STEM data from a LSMO system with embedded NiO nano islands as an example (Fig. 6a). Here, a first ensemble of models is trained using a generic training set from a simple cubic lattice with random atomic displacements. When applied to real experimental data its

Fig. 6 Application of ELIT to a single image. **a** Experimental STEM image of NiO-LSMO. **b** Result (semantically segmented map) from ensemble of models trained on simulated data. **c** Prediction (semantically segmented map) after the first ELIT re-training. **d, e** Predictions (semantically segmented maps) after the second ELIT retraining for a single class (**d**) and multiple classes (**e**).

models mostly fail to identify the second (NiO) structure and tend to ‘find’ some unphysical features (Fig. 6b). Here we first used the best ensemble model to retrain an ensemble of models on the same simulated data using the weights of the best model as a baseline. This allowed for uncertainty estimation in the model’s prediction that was used to select only the robust atomic features in the prediction, i.e., features associated with high variance in uncertainty maps were filtered out. The atomic positions associated with the remaining low-variance features were used to construct a new training set from the experimental data, which provided a significant improvement in the quality and rate of atomic segmentation/detection (Fig. 6c). Next, one may continue re-training a single class model until achieving a nearly perfect detection rate (Fig. 6d) or switch to a multi-class classification as described earlier to categorize multiple sublattices in the system (Fig. 6e).

In summary, we have developed the ELIT framework for the iterative identification of features and corresponding pixel-wise uncertainty maps from high-resolution STEM images and demonstrated its application for several systems. We found that when applying model(s) trained on simulated data to real experimental data the EL can be used for selecting a subset of ‘physical’ (artifact-free) models. These models can be iteratively retrained on ‘good’ portions of experimental data in which case the new ensembles can be used for improving the robustness of predictions and providing meaningful uncertainty estimations on the level of individual pixels. The method can be extended to other than U-Net deep learning architectures as well as to different imaging techniques.

This approach allows for two significant advances in DL applications to experiments. The first is the potential of ensemble-based uncertainty quantification. For the cases explored here and in the author’s experience, uncertainties are often associated with unusual phenomena and can serve as a flag for unexpected behaviors. Perhaps even more importantly, this approach enables rapid correction for out of

distribution drift during automated experiments where the imaging conditions change compared to the training set. Namely, it enables substituting the slow labeling/retraining of the network to rapid human-based deselection of the network from an ensemble.

METHODS

Samples preparation

Atmospheric pressure chemical vapor deposition (AP-CVD) was used to grow graphene on Cu foil⁵⁴. A coating of poly(methyl methacrylate) (PMMA) was spin coated over the surface to protect the graphene and form a mechanical stabilizer during handling. Ammonium persulfate dissolved in deionized (DI) water was used to etch away the Cu foil. The remaining PMMA/graphene stack was rinsed in DI water and positioned on a TEM grid and baked on a hot plate at 150 °C for 15 min to promote adhesion between the graphene and TEM grid. After cooling, acetone was used to remove the PMMA and isopropyl alcohol was used to remove the acetone residue. The sample was dried in air and baked in an Ar-O₂ atmosphere (10% O₂) at 500 °C for 1.5 h to remove residual contamination⁵⁵. Before examination in the STEM, the sample was baked in vacuum at 160 °C for 8 h.

The LSMO-NiO VAN and the single-phase LSMO and NiO films were grown on STO (001) single-crystal substrates by PLD using a KrF excimer laser ($\lambda = 248$ nm) with fluence of 2 J/cm² and a repetition rate of 5 Hz. All films were grown at 200 mTorr O₂ and 700 °C. The films were post-annealed in 200 Torr of O₂ at 700 °C to ensure full oxidation, and cooled down to room temperature at a cooling rate of 20 °C/min. For out-of-plane transport measurements, the films were grown on 0.5% Nb-doped STO (001) single-crystal substrates. The film composition was varied by using composite laser ablation targets with different compositions.

STEM imaging

The plan-view STEM samples of NiO-LSMO were prepared using ion milling after mechanical thinning and precision polishing. In brief, a thin film sample was firstly ground, and then dimpled and polished to a thickness <20 μm from the substrate side. The sample was then transferred to an ion milling chamber for further substrate-side thinning.

The ion beam energy and milling angle were adjusted towards lower values during the thinning process, which was stopped when an open hole appeared for STEM characterization. The STEM used for the characterization was a Nion UltraSTEM200 operated at 200 kV. The beam illumination half-angle was 30 mrad and the inner detector half-angle was 65 mrad. Electron energy-loss spectra were obtained with a collection half-angle of 48 mrad.

For graphene imaging, a Nion UltraSTEM 200 was used, operated at 100 kV accelerating voltage with a nominal beam current of 20 pA and nominal convergence angle of 30 mrad. Images were acquired using the high-angle annular dark-field detector.

Ab initio molecular dynamics (AIMD)

Details on AIMD simulations for the graphene movies: Ab initio quantum-mechanical MD simulations were performed using the projector augmented plane-wave (PAW) method and PAW-PBE potential⁵⁶ as implemented in the Vienna ab initio simulation package (VASP)^{57,58}.

A graphene supercell of 199 atoms with lattice parameters $a = b = 24.68 \text{ \AA}$, $c = 8.60 \text{ \AA}$ with $\alpha = \beta = 90^\circ$, $\gamma = 59.99^\circ$ was considered for performing the simulations. All computations were carried out with a 400-eV plane-wave cutoff energy with appropriate Monkhorst Pack⁵⁹ k-point meshes, at 300 K temperature with 2000 timesteps of 1 fs each. Atoms were randomly removed from the supercell one-by-one and the MD simulation was repeated until 10 atoms remained in the supercell. The final converged coordinates for every timestep as produced by all the simulations were then used to prepare training set for building ensemble of models.

For NiO-LSMO, a generic square lattice with random displacements was utilized to simulate the trajectories, and used for training an ensemble of models in the first step.

DATA AVAILABILITY

The data used for all analysis are available through the Jupyter notebooks located at <https://github.com/aghosh92/ELIT>.

CODE AVAILABILITY

All the deep learning routines were implemented using a home-built open-source software package AtomAI (<https://github.com/pycrosopy/atomai>) and all data are readily available via an interactive Jupyter notebook at <https://github.com/aghosh92/ELIT>.

Received: 12 February 2021; Accepted: 10 June 2021;

Published online: 02 July 2021

REFERENCES

- Pennycook, S. J. & Nellist, P. D. (eds) *Scanning Transmission Electron Microscopy: Imaging and Analysis* (Springer, 2011).
- Krivanek, O. L. et al. Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy. *Nature* **464**, 571–574 (2010).
- Gerber, C. & Lang, H. P. How the doors to the nanoworld were opened. *Nat. Nanotechnol.* **1**, 3–5 (2006).
- Mannhart, J. & Schlom, D. G. Oxide interfaces—an opportunity for electronics. *Science* **327**, 1607–1611 (2010).
- Moler, K. A. Imaging quantum materials. *Nat. Mater.* **16**, 1049–1052 (2017).
- Keimer, B. & Moore, J. E. The physics of quantum materials. *Nat. Phys.* **13**, 1045–1055 (2017).
- Jia, C. L., Urban, K. W., Alexe, M., Hesse, D. & Vrejoiu, I. Direct observation of continuous electric dipole rotation in flux-closure domains in ferroelectric Pb(Zr, Ti)O₃. *Science* **331**, 1420–1423 (2011).
- Jia, C. L. et al. Unit-cell scale mapping of ferroelectricity and tetragonality in epitaxial ultrathin ferroelectric films. *Nat. Mater.* **6**, 64–69 (2007).
- Chisholm, M. F., Luo, W. D., Oxley, M. P., Pantelides, S. T. & Lee, H. N. Atomic-scale compensation phenomena at polar interfaces. *Phys. Rev. Lett.* **105**, 197602 (2010).
- Borisevich, A. Y. et al. Interface dipole between two metallic oxides caused by localized oxygen vacancies. *Phys. Rev. B* **86**, 140102 (2012).
- Yadav, A. K. et al. Observation of polar vortices in oxide superlattices. *Nature* **530**, 198–201 (2016).
- Nelson, C. T. et al. Spontaneous vortex nanodomain arrays at ferroelectric heterointerfaces. *Nano Lett.* **11**, 828–834 (2011).

- Jia, C. L. et al. Oxygen octahedron reconstruction in the SrTiO₃/LaAlO₃ heterointerfaces investigated using aberration-corrected ultrahigh-resolution transmission electron microscopy. *Phys. Rev. B* **79**, 081405 (2009).
- Borisevich, A. Y. et al. Mapping octahedral tilts and polarization across a domain wall in BiFeO₃ from Z-contrast scanning transmission electron microscopy image atomic column shape analysis. *ACS Nano* **4**, 6071–6079 (2010).
- Borisevich, A. Y. et al. Suppression of octahedral tilts and associated changes in electronic properties at epitaxial oxide heterostructure interfaces. *Phys. Rev. Lett.* **105**, 087204 (2010).
- He, Q. et al. Towards 3D mapping of BO₆ octahedron rotations at perovskite heterointerfaces, unit cell by unit cell. *ACS Nano* **9**, 8412–8419 (2015).
- Kim, Y. M. et al. Probing oxygen vacancy concentration and homogeneity in solid-oxide fuel-cell cathode materials on the subunit-cell level. *Nat. Mater.* **11**, 888–894 (2012).
- Kim, Y. M. et al. Direct observation of ferroelectric field effect and vacancy-controlled screening at the BiFeO₃/LaSr_{1-x}MnO₃ interface. *Nat. Mater.* **13**, 1019–1025 (2014).
- Lin, W. Z. et al. Local crystallography analysis for atomically resolved scanning tunneling microscopy images. *Nanotechnology* **24**, 415707 (2013).
- Sang, X. H. et al. In situ edge engineering in two-dimensional transition metal dichalcogenides. *Nat. Commun.* **9**, 1–7 (2018).
- Sang, X. H. et al. In situ atomistic insight into the growth mechanisms of single layer 2D transition metal carbides. *Nat. Commun.* **9**, 1–9 (2018).
- Taheri, M. L. et al. Current status and future directions for in situ transmission electron microscopy. *Ultramicroscopy* **170**, 86–95 (2016).
- Mehdi, B. L. et al. In-situ electrochemical transmission electron microscopy for battery research. *Microsc. Microanal.* **20**, 484–492 (2014).
- Abellan, P. et al. Probing the degradation mechanisms in electrolyte solutions for Li-ion batteries by in situ transmission electron microscopy. *Nano Lett.* **14**, 1293–1299 (2014).
- Ishikawa, R. et al. Direct observation of dopant atom diffusion in a bulk semiconductor crystal enhanced by a large size mismatch. *Phys. Rev. Lett.* **113**, 155501 (2014).
- Yang, Z. Q. et al. Direct observation of atomic dynamics and silicon doping at a topological defect in graphene. *Angew. Chem. -Int. Ed.* **53**, 8908–8912 (2014).
- Susi, T., Mayer, J. C. & Kotakoski, J. Manipulating low-dimensional materials down to the level of single atoms with electron irradiation. *Ultramicroscopy* **180**, 163–172 (2017).
- Dyck, O., Kim, S., Kalinin, S. V. & Jesse, S. Placing single atoms in graphene with a scanning transmission electron microscope. *Appl. Phys. Lett.* **111**, 113104 (2017).
- Mishra, R., Ishikawa, R., Lupini, A. R. & Pennycook, S. J. Single-atom dynamics in scanning transmission electron microscopy. *MRS Bull.* **42**, 644–652 (2017).
- Maksov, A. et al. Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS₂. *Npj Comput. Mater.* **5**, 1–8 (2019).
- Krull, A., Hirsch, P., Rother, C., Schiffrin, A. & Krull, C. Artificial-intelligence-driven scanning probe microscopy. *Commun. Phys.* **3**, 1–8 (2020).
- Kalinin, S. V., Sumper, B. G. & Archibald, R. K. Big-deep-smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
- Kalinin, S. V. et al. Big, deep, and smart data in scanning probe microscopy. *ACS Nano* **10**, 9068–9086 (2016).
- Kalinin, S. V., Borisevich, A. & Jesse, S. Fire up the atom forge. *Nature* **539**, 485–487 (2016).
- Dyck, O., Jesse, S. & Kalinin, S. V. A self-driving microscope and the Atomic Forge. *MRS Bull.* **44**, 669–670 (2019).
- Dyck, O. et al. Building structures atom by atom via electron beam manipulation. *Small* **14**, 1801771 (2018).
- Jesse, S. et al. Direct atomic fabrication and dopant positioning in Si using electron beams with active real-time image-based feedback. *Nanotechnology* **29**, 255303 (2018).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2017).
- He, K., Zhang, X., Ren, S. & Sun, J. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, <https://doi.org/10.1109/CVPR.2016.90>.
- Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Med. Image Comput. Assist. Interv.* **9351**, 234–241 (2015).
- Ziatdinov, M., Maksov, A. & Kalinin, S. V. Learning surface molecular structures via machine vision. *Npj Comput. Mater.* **3**, 1–9 (2017).
- Rashidi, M. & Wolkow, R. A. Autonomous scanning probe microscopy in situ tip conditioning through machine learning. *ACS Nano* **12**, 5185–5189 (2018).
- Ziatdinov, M. et al. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **11**, 12742–12752 (2017).

44. Gordon, O. M., Hodgkinson, J. E. A., Farley, S. M., Hunsicker, E. L. & Moriarty, P. J. Automated searching and identification of self-organized nanostructures. *Nano Lett.* **20**, 7688–7693 (2020).
45. Horwath, J. P., Zakharov, D. N., Mégret, R. & Stach, E. A. Understanding important features of deep learning models for segmentation of high-resolution transmission electron microscopy images. *Npj Comput. Mater.* **6**, 1–9 (2020).
46. Lee, C.-H. et al. Deep learning enabled strain mapping of single-atom defects in two-dimensional transition metal dichalcogenides with sub-picometer precision. *Nano Lett.* **20**, 3369–3377 (2020).
47. Ovadia, Y. et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Adv. Neural Inf. Process Syst.* **33**, 13991–14002 (2019).
48. Zech, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* **15**, e1002683 (2018).
49. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process Syst.* **31**, 6405–6416 (2017).
50. Wilson, A. G. & Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. Preprint at <https://arxiv.org/pdf/2002.08791.pdf> (2020).
51. Fort, S., Hu, H. & Lakshminarayanan, B. Deep Ensembles: A Loss Landscape Perspective. Preprint at <https://arxiv.org/pdf/1912.02757.pdf> (2019).
52. Izmailov, P., Wilson, A. G., Podoprikin, D., Vetrov, D. & Garipov, T. Averaging weights leads to wider optima and better generalization. *UAI* **2**, 876–885 (2018).
53. Barthel, J. Dr. Probe: A software for high-resolution STEM image simulation. *Ultramicroscopy* **193**, 1–11 (2018).
54. Vlasiouk, I. et al. Large scale atmospheric pressure chemical vapor deposition of graphene. *Carbon* **54**, 58–67 (2013).
55. Dyck, O., Kim, S., Kalinin, S. V. & Jesse, S. Mitigating e-beam-induced hydrocarbon deposition on graphene for atomic-scale scanning transmission electron microscopy studies. *J. Vac. Sci. Technol. B* **36**, 011801 (2017).
56. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
57. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
58. Kreese, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (1993).
59. Monkhorst, H. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188 (1976).

ACKNOWLEDGEMENTS

This effort (machine learning) is based upon work supported by the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences Data, Artificial Intelligence and Machine Learning at DOE Scientific User Facilities (A.G., S.V.K., B.G.S.) and was also supported (STEM experiment) by the DOE, Office of Science, Basic Energy Sciences (BES), Materials Sciences and Engineering Division (O.D.), and was performed and partially supported (M.Z., B.G.S.) at the Oak Ridge National Laboratory's Center for Nanophase Materials Sciences (CNMS), a DOE Office of

Science User Facility. Dr. Matthew Chisholm (ORNL) is gratefully acknowledged for the STEM data on Ni-LSMO used in this work.

AUTHOR CONTRIBUTIONS

S.V.K. and M.Z. conceived the project. A.G. prepared training data from simulations, trained ELIT models, and implemented all the associated workflows in Jupyter notebooks. A.G. and B.G.S. ran the MD simulations. O.D. obtained STEM data on graphene. M.Z. realized the ELIT framework via AtomAI and PyTorch libraries. M.Z., S.V.K., and A.G. wrote the drafts of the manuscript while B.G.S. and O.D. contributed to the writing.

COMPETING INTERESTS

S.V.K. is a member of the Editorial Board for npj Computational Materials. S.V.K. was not involved in the journal's review of, or decisions related to, this manuscript.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00569-7>.

Correspondence and requests for materials should be addressed to S.V.K. or M.Z.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021