npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Classifying handedness in chiral nanomaterials using label error robust deep learning

C. K. Groschner[1], Alexander J. Pattison[2], Assaf Ben-Moshe[3,4], A. Paul Alivisatos[3,4], Wolfgang Theis[2] and M. C. Scott[1,5]✉

High-throughput scanning electron microscopy (SEM) coupled with classification using neural networks is an ideal method to determine the morphological handedness of large populations of chiral nanoparticles. Automated labeling removes the time-consuming manual labeling of training data, but introduces label error, and subsequently classification error in the trained neural network. Here, we evaluate methods to minimize classification error when training from automated labels of SEM datasets of chiral Tellurium nanoparticles. Using the mirror relationship between images of opposite handed particles, we artificially create populations of varying label error. We analyze the impact of label error rate and training method on the classification error of neural networks on an ideal dataset and on a practical dataset. Of the three training methods considered, we find that a pretraining approach yields the most accurate results across label error rates on ideal datasets, where size and other morphological variables are held constant, but that a co-teaching approach performs the best in practical application.

npj Computational Materials (2022)8:149 ; https://doi.org/10.1038/s41524-022-00822-7

INTRODUCTION

There is growing interest in inorganic chiral nanomaterials for application in optoelectronics and biomimetics[1–3]. Specific parameters during wet chemical synthesis of chiral nanomaterials[4–8] can induce a large degree of structural variety. Of particular importance, synthesis parameters can favor one handedness over another. For example, despite their underlying chiral crystal structure, Tellurium (Te) nanoparticles can have different ratios of certain chiralities depending on synthesis conditions[4,5]. In another example from recent work by van der Boom et al., tuning of synthesis parameters gave rise to metallo-organic single crystals with a wide variety of complex non-trivial morphologies, many of which are chiral[9,10]. These variations are induced by many factors including thermodynamic versus kinetic growth pathways, and differences in interactions of chiral organic molecules with small clusters of atoms during synthesis. Therefore, precise tuning of chirality alongside size via wet chemical synthesis is yet to be obtained in many systems. To be able to tune chirality, one must first determine the influence of the many synthetic parameters, such as temperature, precursor concentration, or concentration and type of structure-directing chiral ligands, on the outcome population. This need for high-throughput analysis motivates the development of methods to classify handedness in chiral nanoparticle populations with the goal of determining the influence of these synthetic parameters.

While circular dichroism (CD) measurements are sensitive to chirality in Te nanoparticles, it is very challenging to extract quantitative information about the abundance of each handedness, as the molar CD of these materials is unknown and can only be estimated[4,5]. Scanning electron microscopy (SEM), in contrast, can be used to unambiguously determine the handedness of morphologically chiral nanomaterials[11–13]. SEM is sensitive to surface topology and can directly determine morphological chirality and handedness, unlike (scanning) transmission electron microscopy ((S)TEM) methods that sum information along the beam direction such that faceting information can be lost and therefore requires that multiple images be used to determine handedness[14]. High-throughput SEM imaging is therefore a particularly promising way to measure the size and handedness of large populations of chiral nanoparticles to better understand the role of synthetic variables on outcome populations. However, determining particle statistics by hand from high-throughput data is extremely laborious and time consuming. Due to the increasing ease of implementing neural networks for image analysis[15–17], deep learning is a promising replacement for manual analysis. Yet, deep learning is known for requiring large training datasets, which still need manual labeling, meaning that the application of deep learning to chirality studies could also be prohibitively time consuming due to expert labeling requirements.

Given that synthesis routes yielding chiral materials often favor one handedness over the other, we have found that it is possible to label the handedness of all the particles in the dataset by first labeling them with the dominant handedness, and then mirroring these images to create a dataset labeled with the opposite handedness. This labeling process is demonstrated in Fig. 1. Labeling all images with the majority handedness of course leads to a dataset with a specific fraction of erroneous labels that is equal to the fraction of particles that did not have the dominant handedness. For synthesis conditions that yield almost exclusively one handedness, this automated mirror labeling strategy is very successful, but for other conditions, it can yield a significant number of mislabeled images. The question then becomes how to extend this automated labeling method across synthesis conditions, such that the accuracy of networks is not hampered by mislabeled data.

Generating accurate models from erroneously labeled datasets, also known as noisy datasets in the machine learning community, has been an expanding area of research. Deep learning is known to be able to memorize random inputs during training[18], and thus has the potential to memorize random label errors. What makes

[1]Department of Materials Science and Engineering, UC Berkeley, Berkeley, CA, USA. [2]School of Physics and Astronomy, University of Birmingham, Birmingham, UK. [3]Department of Chemistry, University of California, Berkeley, Berkeley, CA, USA. [4]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, Berkeley, CA, USA. ✉email: mary.scott@berkeley.edu

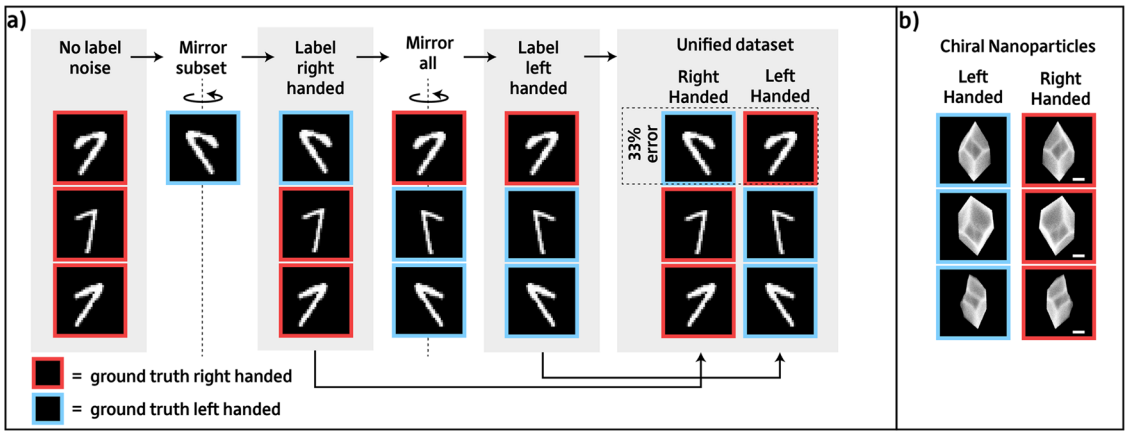

Fig. 1 The mirror labeling scheme used for data labeling with sample images. a Schematic demonstrating the mirror labeling procedure by which training datasets are automatically labeled and the error rate in the dataset is controlled on the MNIST seven dataset. **b** Examples of images for both classes from the chiral nanoparticle datasets. Scale bars, 100 nm.

true learning, and thus generalizability, possible is that before memorizing real data, repeated motifs within the data are learned by the network[19]. Networks can therefore recognize predictive features without directly memorizing inputs. Deep neural networks are thus able to learn the true signal from datasets with 100:1 erroneous to true labels[20]. To further extend the label error tolerance of deep neural networks several strategies have been implemented. These methods can be broadly separated into four types of approaches: (1) label error robust techniques[20–22], (2) label error reduction methods[23,24], (3) label error estimation[25,26], and (4) preprocessing pipelines[27]. This paper will focus on label error robust and label error reduction strategies. Previous research into these methods focuses on how these methods perform on very large datasets of real images with a large number of classes. Microscopy data, in contrast, contains fewer classes but also much smaller datasets than standard computer vision datasets.

To test the influence of label error in a controlled manner, we create an idealized dataset where only the handedness of the particle is correlated with label error. To create the idealized dataset we take a manually, accurately labeled dataset of all right-handed particles and mirror a certain fraction of particles to create the specified label error fraction as is described in Fig. 1, so that we can model different possible synthesis outcomes. We use these idealized datasets to test the performance of several neural networks when trained with the various label error rates. To test whether the networks will perform in the same way with real label error, we then also test the neural networks on as-synthesized micrographs of left and right-handed particles. We then examine the three neural networks' performance on the idealized and real label error datasets.

In this work, we focus on label error robust and label error-reducing methods that do not require the creation of a large labeled dataset or other supervision. First, we examine the performance of a standard convolutional neural network (CNN). The architecture for the network is shown in Fig. 1a. It consists of two convolutional and max-pooling blocks with a final dropout and dense unit. We compare this standard CNN to the co-teaching training procedure developed by Han et al.[24]. The co-teaching method uses the difference in loss between two networks trained in tandem on the error-containing dataset to remove incorrectly labeled data. The cleaned dataset is then used to train its partner model. This process of cleaning and trading datasets is then repeated. For consistency, the two tandem networks used in our study of co-teaching consist of the same architecture as the standard CNN we tested, as shown in Fig. 1a. A schematic of the co-teaching training procedure is provided in Fig. 1c. Finally, we propose a new method inspired by label error estimation and

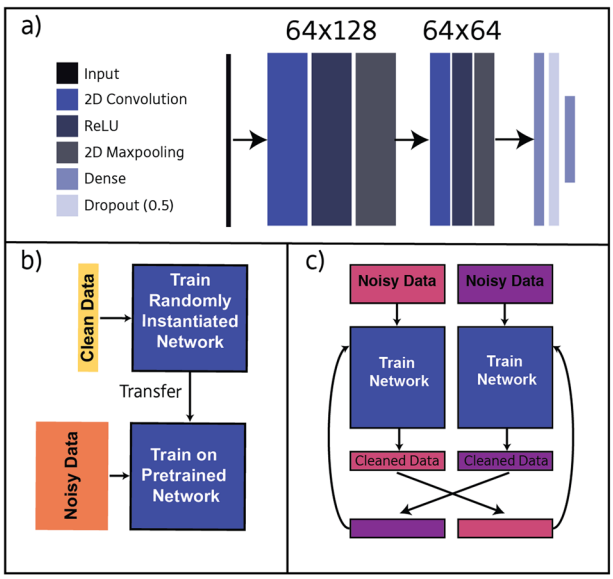

Fig. 2 Schematics for the three methods studied. All networks use the architecture shown in part a. **a** Schematic of convolutional neural network architecture implemented with features by height and width labeled. **b** Schematic of the two-step training procedure. Randomly instantiated network is trained for one training round, i.e., epoch, using error-free data before network weights are transferred and trained using the dataset with label error. **c** Schematic of co-teaching training procedure. Two networks are trained in parallel and after each training round the labels judged as erroneous are removed by only passing data with low enough loss to the twin network.

curriculum learning[28–31]. In this two-step training method, we use a very small label error-free dataset to start network learning before training on the dataset with label error. The neural network architecture is the same as the standard CNN shown in Fig. 2a and the training procedure is illustrated in Fig. 2b. This new two-step method leverages the memorization effect, in order to push networks toward learning relevant features even at high error rates[32] while avoiding data loss by removing incorrect labels.

From these experiments, we show that, for the idealized datasets, the label error-reducing and co-teaching strategies achieve similar accuracy rates, but that co-teaching method leads to the most consistent results across label error rates. We find that if label error is correlated with variations in particle appearance

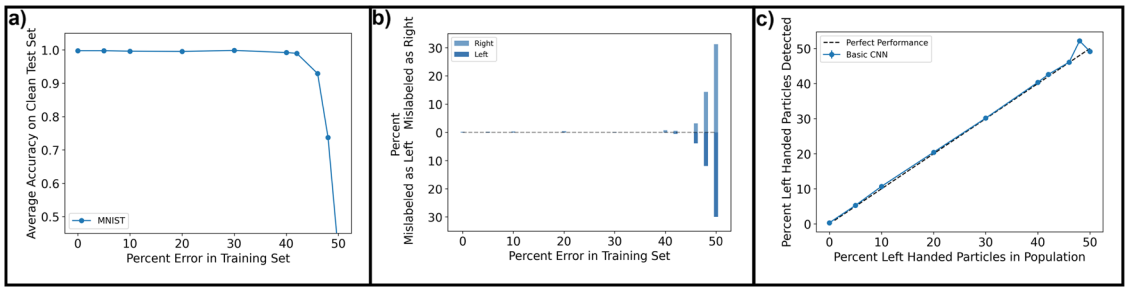

Fig. 3 Figures quantifying the performance of the standard CNN on the MNIST training sets. a Accuracy of standard CNN on the MNIST seven error-free ground truth test set given increasing error in the training set. **b** The percent of sevens erroneously labeled as left or right handed in the test set. **c** Percent of left-handed sevens in a population predicted by a network that was trained with data containing that percent of lefts. Error bars are obtained by training each architecture multiple times and then calculating the standard error on the accuracy of the test set by the multiple versions of the network.

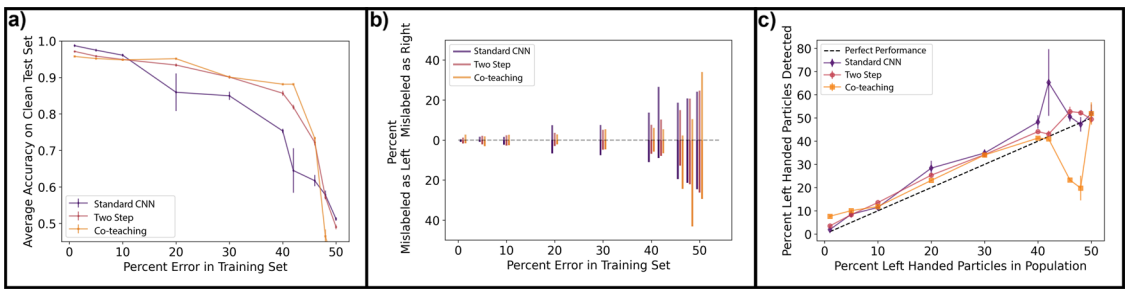

Fig. 4 Figures quantifying the performance of the three methods on the Te training sets. a Average accuracy of networks from each technique on the error-free ground truth test set of Te SEM data given increasing label error in the Te training set. **b** The percent of right and left-handed particles that are misclassified by each network. **c** Predicted fraction of left-handed Te particles for a population with a specified percent of left-handed particles using the network that corresponds to training with that percent of left-handed particles. Error bars are obtained by training each architecture multiple times and then calculating the standard error on the accuracy of the test set by the multiple versions of the network.

besides chirality, as is found in the as-synthesized dataset, only the co-teaching method maintains high accuracy. We hope the exploration of these methods will enable the wider application of weak labeling and therefore deep learning for the chiral materials community and the materials community more broadly.

RESULTS

Analysis methods

For each technique, we assessed the network's performance on an error-free ground truth test set with 50% left and right-handed images after being trained on datasets with different label error fractions. We also analyzed the distribution of error between the two classes by plotting the histogram of incorrectly labeled left-handed particles and a number of incorrectly labeled right-handed particles in the test dataset. Dramatically misclassifying one but not the other label could lead to a significant skew in the calculated ratio between classes, a metric that is very important in studying chirality. We finally evaluate the ability of the networks to recover the fraction of particles belonging to the minority population, to simulate the use of these methods for data from real synthesis procedures. To analyze the ability to accurately recover the minority population, we use a test set that has the same mislabel fraction as the training set. Using this second test set, we plot $f(x) = \frac{n_{left}(x)}{n_{total}}$, where x is the mislabeled fraction of the training set, $n_{\text{left}}(x)$ is the total number of particles labeled left handed by the network for that test set, and n_{total} is the total number of particles in the test set. All networks were trained on datasets containing label error. The label error generation process is outlined in the Methods section and illustrated in Fig. 1a. For each method, we trained three separate networks using that method and plot the average performance metric and the standard error.

MNIST sevens

Standard convolutional neural network. Before testing our three label error robust neural network approaches, we first tested a standard convolutional neural net's ability to learn to classify chiral images in the presence of varying amounts of label error using images of the number seven from the MNIST dataset. The training set for the model contained 10,024 images and the test set 1258 images. The architecture used is shown in Fig. 2a. Figure 3a shows that a standard CNN can learn the classes accurately up to 42% label error. Figure 3b shows the percent of misclassified particles that were left handed vs right handed demonstrating that there is minimal bias, in terms of handedness, in the classification error as well. Figure 3c shows that, given an enantiomerically skewed population, the neural network trained with the corresponding amount of label error is able to correctly predict the fraction of sevens that have that minority handedness. All these metrics suggest that the mirror labeling procedure works and even a standard CNN can handle the label error presented.

Application to idealized datasets of Te chiral nanoparticles

Standard convolutional neural network. We then applied the same mirror labeling and training protocol from the MNIST dataset to the SEM images of chiral Te nanoparticles. Sample images of the chiral particles are shown in Fig. 1b and in Supplementary Figs. 1–3. The training set for the chiral nanoparticle models contained 3062 images, before augmentation, and the test set 382 images. Further information on augmentation is presented in the Methods section. Although it is possible there are structural differences between the left and right-handed populations, the data were

Table 1. The average accuracy and standard error on the error-free, as-synthesized balanced test set for each of the three techniques.

	Average accuracy as-synthesized
Standard CNN	0.65 ± 0.02
Two-step training	0.67 ± 0.03
Co-teaching	0.915 ± 0.003

The as-synthesized test set contains an equal number of left- and right-handed particle images.

augmented to ensure that handedness was not correlated with arbitrary image features such as illumination. For the standard CNN on SEM images, the error-free test set accuracy begins to degrade after 10% label error, as can be seen in Fig. 4a. Reasonable accuracy is still recovered up to 30% label error. In order to examine the maximum amount of label error that can be handled, we increased our sampling close to 50% flip error—close to random labeling. Figure 4b gives the percent of misclassified particles that were left handed vs right handed. This shows that the network on average is rather inconsistent in whether it misclassifies one handedness more than the other at high label error. Figure 4c shows the ability of the network, for an enantiomerically skewed population, to recover the correct fraction of the mislabeled portion of the population. The standard CNN is able to accurately predict the fraction of the minority population (and therefore mislabeled in the training set) up to 40% label error.

Accuracy results for different label error rates using the co-teaching training method are shown in Fig. 4a. We see that at very low label error the standard CNN achieves higher accuracy than co-teaching. However, the co-teaching method is able to achieve much better accuracy than the standard CNN when the label error increases past 10%. One benefit of the co-teaching network is that from 0% to approximately 40% label error the recovered accuracy is very consistent. One promise of the co-teaching network is that datasets with large amounts of label error can still be used for accurate training. However, in Fig. 4a we see that there are still limits on the amount of label error that can be tolerated. We still observe a dramatic reduction in accuracy beyond 40% label error. As shown in Fig. 4b, the co-teaching network does not misclassify one handedness more than the other at low label error rates, but at label error rates above 42% the network appears to be significantly over-predicting left-handed particles. This skew in misclassification is also reflected in the fraction of left-handed particles detected. Figure 4c shows that the co-teaching networks on average, badly underestimate the number of minority handed particles past 42% label error.

Similar to co-teaching two-step training creates a network that performs well up to about 40% label error. Co-teaching and the standard CNN sometimes slightly outperform the two-step method in terms of accuracy but the two-step training creates perhaps the most consistent results. This is reflected in the fact that the two-step training procedure leads to the most consistent detection of minority particles. Unlike the other methods, it does not drastically over- or underestimate the population of left-handed particles at any point, as seen in Fig. 4c. This is a consequence of the fact that the training procedure does not lead to a high bias against one class or the other at any training label error, as can be seen in Fig. 4b.

Application to as-synthesized particles

To assess performance on real data, we applied these methods to a dataset of as-synthesized Te particles. The difference between the as-synthesized and ideal case is that in the ideal case we start with all right-handed particles so all "left-handed" particles are mirror images of right-handed particles. The as-synthesized dataset is made up of particle images from a sample that contained 22% left-handed and 78% right-handed particles. The preparation of these particles can be found elsewhere[5]. We created a training set of 2461 particle images. The mirror labeling procedure was then used without the error creation step, since left-handed particles are now part of the sample. The results are shown in Table 1. The performance of the standard CNN and two-step training procedure does not match that obtained with the ideal dataset. We see that the standard CNN and two-step training procedure fall short of acceptable accuracy. Only the co-teaching method maintains high accuracy.

DISCUSSION

Performance comparison between MNIST and nanoparticle-trained neural nets

A comparison of the accuracy on the error-free test set indicates that the application of a CNN to the MNIST data results in higher accuracy across most error levels than the results from training with the chiral nanoparticle dataset. The accuracy of the network compared to the chiral nanoparticle dataset suggests that the network is fitting a simpler feature space than the chiral nanoparticle case. It should be noted that the MNIST input image is much smaller than the Te, meaning that by keeping the network size the same we have more convolutional kernels relative to the number of pixels in the MNIST network. In addition, to the human observer, it is easy to see that there are only a few prototypical examples for a number seven in the MNIST dataset and few deviations from these prototypes. By inspecting data in Fig. 1a, b, it is clear that while most parts of the seven images contribute to the chiral shape, the chiral particles have many features which are unrelated to the chirality of the particle. In fact, only the center facets of the chiral particles determine a particle's handedness. Therefore, since we have held all hyperparameters constant while training on each dataset, it is reasonable to interpret at least part of the difference in accuracy is due to the difference in richness of the feature space. Furthermore, learning a more complex feature space with fewer features actually contributing to the classification task is likely the reason for the lower accuracy of CNN trained on the chiral nanoparticle dataset.

The limited feature space of the MNIST dataset vs the chiral nanoparticles was by design, for the purpose of comparing label error robustness of the network when applied to datasets with and without "hard examples", those samples which are harder for the network to learn. This was important to consider since previous work has found evidence that CNNs treat hard examples and label error in a similar fashion[18,33]. By proving the label error robust nature of our proposed standard CNN on the MNIST dataset, we could isolate any challenges when applying the same mirror labeling scheme to actual chiral materials which have many more possible representations. We hypothesize that this increased feature space is responsible for the lower performance of the networks on chiral nanoparticle data, particularly the standard CNN, in the presence of high label error compared to the network trained on MNIST. This comparison highlights that label label error robust properties of standard neural networks are heavily dependent on the richness of the feature space.

Comparison of label error robust architectures and training procedures on ideal SEM data

Our results on the ideal SEM dataset suggest that the best method to employ depends on label error and the type of error tolerated. In general two-step training and co-teaching show enhanced robustness to label error, as they achieve a high classification accuracy, until approximately 40% label label error and above.

The standard CNN does not demonstrate the same label error robustness as the other two methods. This deviates from the results of the standard CNN on the MNIST sevens and suggests that for the more complex feature space presented by the Te dataset, the label error robustness of a standard CNN is not as consistent.

Co-teaching provides a way to explore label error-reducing methods. For very low label error the co-teaching achieves the lowest accuracy, which is likely connected to throwing away hard examples during training[18,33]. However, as training error increases co-teaching becomes the most accurate training method. However, once training label error is too high the accuracy of the co-teaching method rapidly deteriorates and shows strong skew in which handedness is misclassified, unlike the two-step method that declines in accuracy but does not show an increase in misclassification skew. It is likely that since the co-teaching method includes no prior knowledge about left and right-handed representations, past a certain label error rate it is no longer able to distinguish the label error from variations in handedness, therefore causing its performance to plummet.

The two-step method achieves almost the same level of accuracy as the standard CNN at low training label error but is also able to sustain high levels of accuracy with more than 40% label error like the co-teaching method. This improved performance at high label error is interesting because we do not use this data to infer any extra information or weighting schemes when training on data with label error unlike some other pretraining methods[28,31]. To our knowledge, other work in this field has used small supervised training sets to train label error estimation layers and other corrective measures but not as a transfer learning procedure to initialize the network. The results we present, therefore, suggest that benefits can also be reaped by utilizing curated, tiny, manually labeled datasets to help the network learn the correct relationships between key features. We hypothesize that the two-step method is exploiting the fact that networks are prone, in early training rounds, to learn simple features as opposed to memorizing data[19,32]. While we rely on the early training rounds learning easy representations from the tiny dataset to direct learning class features correctly, it is likely that by initializing the network with a very small dataset, we are limiting representations the network is sensitive to. This points to an inherent tradeoff in the two-step method between labeling requirements and generalization, which should be explored in future work. Limiting the initial representations learned may also explain why we achieve slightly lower accuracy when training label error is low.

The difference in the misclassification bias between the co-teaching and two-step method is what truly distinguishes these two techniques under the presented conditions. Previous work has shown that pretraining a network with an unsupervised dataset leads to better generalization due to the pretraining acting as regularizer[34]. Though we use a supervised training set for pretraining, the difference in misclassification bias between the two networks supports that particularly at high levels of training label error, the first training round is constraining the network to more balanced learning. This observation is a key finding for the application proposed since being able to consistently recover handedness ratios is vital for understanding the influence of synthesis parameters on the development material handedness.

Comparison of ideal to as-synthesized dataset

The difference in the performance of the three methods on the ideal dataset versus the as-synthesized dataset highlights the importance of the alternate features present between the as-synthesized left and right-handed particles. Analyzing the as-synthesized images (shown in the Supplementary Fig. 3), it is evident that the growth that led to minority left-handed particles in the sample also caused those left-handed particles to have a smaller chiral facet. The particles are also smaller overall. Therefore, in the as-synthesized case, we not only have a complex feature space to learn, but also image features that are directly correlated with the label error. This feature-correlated label error clearly has implications for the methods explored. The fact that augmentation did not enable learning in the standard CNN and two-step training method suggests that there are other aspects of particle shape and contrast that are correlated. Overall, these results indicate that when label noise is correlated with other feature variations, the co-teaching approach is the best method out of those considered here. The features that distinguish the minority class (and therefore mislabeled data) may in fact contribute to the success of the co-teaching method since these features could be learned and therefore excluded improving the performance. Future work will need to be cognizant of these differences between ideal datasets and their actual application.

Implications for experimental applications

The as-synthesized analysis shows that for most experimental applications the co-teaching method is the best approach. The as-synthesized dataset had 22% minority handedness present corresponding to 22% label error. This is a relatively high label error and suggests that the co-teaching method would be able to handle experimental conditions[5], even if there are structural differences between the left and right-handed populations. If the samples are true enantiomers and therefore only vary by a mirror operation, as in the ideal dataset, then there are two clear use cases that require choosing between these networks for practical applications of the mirror labeling approach in the chiral nanomaterial community. For synthesis routes where one handedness is much more favored than the other, training a standard CNN is most likely to yield the most accurate population statistics because under this condition our mirror labeling technique will yield low label error. For chirality studies where synthesis conditions lead to a more balanced set of left and right-handed particles, and therefore higher label error, the two-step neural network is the best option since this network provides accurate results and little bias in class prediction across populations.

There are several use cases for the three methods developed here. The automated mirror labeling system described can be used not only in the realm of SEM chiral image analysis, but for any imaging method where the chiral structure leads to geometrically related images. Outside of chirality classification, the label-error robust methods developed have a wide range of possible use cases across the microscopy community.

In summary, these results give insight for leveraging automatic labeling systems such as ours on other microscopy datasets. We see automated labeling as a potential avenue for analysis in the chirality community but also among other materials where datasets are too large to realistically be manually labeled. We have shown the tradeoffs between the richness of the feature space and label error robustness of CNNs, which should be taken into account when using these techniques for other studies. We have shown that both label error robust and label error-reducing methods perform well for small datasets. However, we find that both two-step training and co-teaching far outperform the inherent label error robustness of a standard CNN. For practical applications, the co-teaching method far outperforms other methods. These results point to the conditions under which these methods should be applied.

METHODS

Dataset generation

Datasets of known label error (known percentages of mislabeled images) were generated by starting with an error-free dataset of right-handed images. Particles were segmented by hand using MATLAB data labeler software prior to dataset generation. A set percentage (varying from 0 to 50%) of the images was randomly selected and mirrored, as shown in Fig. 1. All images in the dataset were then labeled as having right handedness. Then, all the images were mirrored, creating a dataset in which all images are labeled as left handed. This procedure ensures that the specific fraction would always be facing the opposite direction from its label. All image augmentation for training was performed after this operation. We applied this method to two datasets. The first was the number 7 samples from the Keras implementation of the MNIST dataset. We chose to use the number 7 since it is a chiral shape. The second dataset consisted of right-handed Te chiral nanoparticles that were manually segmented and labeled by an expert. The complete dataset consists of 1914 right-handed particles that are mirrored to create a set of 3828 total left and right-handed nanoparticle images. In all, 80% of the data is used in the training set, 10% in the validation set, and 10% in the test set. During training, the images are augmented using a combination of 180° rotations, 5° rotations, up to 30% zoom of the image, and 10% shearing. The augmentations are used to create a dataset of 91,840 images per epoch (training round) for training Keras networks, and 97,984 images for PyTorch networks. The difference in the number of images is due to differences in augmentation implementation between Keras and PyTorch.

Standard convolutional neural network classifier

The neural networks were developed with Keras. A simple CNN was implemented for both the MNIST and Te nanoparticle datasets; which contained two convolutional residual units, a schematic is shown in Fig. 2a. Each residual unit contained the convolutional layer, a ReLu layer, and a max-pooling layer. After the residual units, the features are flattened, and passed to a dense layer, dropout, and final dense layer. The logits from the dense layer are then passed to a softmax activation function. Training was done with a categorical crossentropy loss function, and adadelta optimizer. The learning rate was 0.001, the batch size was 32, and ran for 100 epochs. Model checkpointing was used so that model weights were only saved if the validation loss had decreased.

Two-step training classifier

The network implemented for two-step training was also developed with Keras and employed the same neural network architecture as the standard CNN. A schematic is shown in Fig. 2b. The training of the network consists of two distinct steps. First, the network is trained on an extremely small set of images with error-free labels. For this first step, we used ten left-handed particle images and ten right-handed particle images, which are augmented to represent 288 samples per epoch. The network is trained on the tiny, error-free training set for ten epochs. The network is then trained in the same way as the standard CNN for 100 epochs.

Co-teaching classifier

Networks were developed using PyTorch. The code and procedure were adapted from work done by Han et al.[24]. The co-teaching method creates two identical CNNs. Two separate batches of data are given on each network for training. The data whose loss is below a certain threshold then get passed to the other network for training and the data above are removed[24]. A schematic of the procedure is shown in Fig. 2c. In this way, incorrect labels are removed, under the assumption that data with incorrect labels will lead to higher loss. The architecture of the two CNNs was changed to match the CNN architecture used in the Keras-based handedness classifier. The only difference was the addition of an average pooling layer to compensate for differences in layer implementations. The models were trained on the same dataset. The learning rate was 0.001, the batch size was 128, and the maximum number of epochs was 40.

Image acquisition

Tellurium nanoparticles suspended in an aqueous solution were dropcast onto a silicon wafer. Micrographs were acquired on FEI Helios G4 UX at 2 kV using a through-lens detectors with a working distance of 2 mm. Images were collected using Maps 2.5 Software, which collected a grid of approximately 1200 images of collections of nanoparticles (approximately 1400-by-900 pixels each). This large-scale acquisition was stitched together using the same Maps 2.5 software. Nanoparticles were manually segmented from the larger images to make the small input images for the neural networks. Sample images are provided in Supplementary Fig. 2.

Synthesis of Te nanoparticles

The synthesis follows a methodology developed by Ben-Moshe et al.[4], with some modifications. 5.5 ml water, 15 mg TeO_2, and 20 µl NaOH (1 M in H_2O) were stirred vigorously in a 20 ml glass vial (at room temperature), before 2.5 ml of hydrazine hydrate (80% solution) were added in one go. Then, 25 s after the addition of hydrazine, 1 ml of a 100 mM solution of D-penicillamine (adjusted to pH 11 using NaOH solution) was added in one go. The reaction was stopped after 3 h, by diluting twice with a 100 mM SDS solution, followed by repeated cycles of cleaning using centrifugation (6000 RPM, 10 min) and dispersion in water.

DATA AVAILABILITY

All image and label data that support the findings of this study are available on Zenodo: https://doi.org/10.5281/zenodo.5009042[35].

CODE AVAILABILITY

All code for this study is available at https://github.com/ScottLabUCB/LabelFreeChirality.

Received: 20 June 2021; Accepted: 3 June 2022;
Published online: 12 July 2022

REFERENCES

1. Ma, W. et al. Attomolar DNA detection with chiral nanorod assemblies. *Nat. Commun.* **4**, 2689 (2013).
2. Wang, P.-p, Yu, S. J. & Ouyang, M. Assembled suprastructures of inorganic chiral nanocrystals and hierarchical chirality. *J. Am. Chem. Soc.* **139**, 6070–6073 (2017).
3. Zhang, H. et al. Engineering of chiral nanomaterials for biomimetic catalysis. *Chem. Sci.* **11**, 12937–12954 (2020).
4. Ben-Moshe, A. et al. Enantioselective control of lattice and shape chirality in inorganic nanostructures using chiral biomolecules. *Nat. Commun.* **5**, 4302 (2014).
5. Ben-Moshe, A. et al. The chain of chirality transfer in tellurium nanocrystals. *Science* **372**, 729–733 (2021).
6. Wang, P.-p, Yu, S.-J., Govorov, A. O. & Ouyang, M. Cooperative expression of atomic chirality in inorganic nanostructures. *Nat. Commun.* **8**, 14312 (2017).
7. Ma, W. et al. Chiral inorganic nanostructures. *Chem. Rev.* **117**, 8041–8093 (2017).
8. Hananel, U., Ben-Moshe, A., Tal, D. & Markovich, G. Enantiomeric control of intrinsically chiral nanocrystals. *Adv. Mater.* **32**, e1905594 (2020).
9. di Gregorio, M. C. et al. Emergence of chirality and structural complexity in single crystals at the molecular and morphological levels. *Nat. Commun.* **11**, 380 (2020).
10. Singh, V. et al. Unusual surface texture, dimensions and morphology variations of chiral and single crystals. *Angew. Chem. Int. Ed.* **60**, 18256–18264 (2021).
11. Kim, H. et al. γ-Glutamylcysteine- and cysteinylglycine-directed growth of chiral gold nanoparticles and their crystallographic analysis. *Angew. Chem. Int. Ed.* **59**, 12976–12983 (2020).
12. Cho, N. H. et al. Cysteine induced chiral morphology in palladium nanoparticle. *Part. Part. Syst. Charact.* **36**, 1–5 (2019).
13. Rafiei Miandashti, A., Khosravi Khorashad, L., Kordesch, M. E., Govorov, A. O. & Richardson, H. H. Experimental and theoretical observation of photothermal chirality in gold nanoparticle helicoids. *ACS Nano* **14**, 4188–4195 (2020).
14. Dong, Z. & Ma, Y. Atomic-level handedness determination of chiral crystals using aberration-corrected scanning transmission electron microscopy. *Nat. Commun.* **11**, 1–6 (2020).
15. Ede, J. M. Deep learning in electron microscopy. *Mach. Learn.: Sci. Technol.* **2**, 011004 (2021).
16. Chollet, F. et al. Keras. https://keras.io (2015).
17. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H. et al.) 8024–8035 (Curran Associates, Inc., 2019).
18. Zhang, C., Recht, B., Bengio, S., Hardt, M. & Vinyals, O. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings* (ICLR, 2017).

19. Arpit, D. et al. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of *Proceedings of Machine Learning Research* (eds Precup, D. & Teh, Y. W.) 233–242 (PMLR, 2017).
20. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. Preprint at https://arxiv.org/abs/1705.10694 (2017).
21. Wang, Y. et al. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV, 2019).
22. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R. & Qu, L. Making deep neural networks robust to label noise: a loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR, 2017).
23. Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G. & Mohd-Yusof, J. Combating label noise in deep learning using abstention. In Chaudhuri, K. & Salakhutdinov, R. (eds) In *Proceedings of the 36th International Conference on Machine Learning,* Vol. 97 *of Proceedings of Machine Learning Research*, 6234–6243 (PMLR, 2019).
24. Han, B. et al. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, Vol. 31 (eds Bengio, S. et al.) (Curran Associates, Inc., 2018).
25. Arazo, E., Ortego, D., Albert, P., O'Connor, N. & Mcguinness, K. Unsupervised label noise modeling and loss correction. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97 of *Proceedings of Machine Learning Research* (eds Chaudhuri, K. & Salakhutdinov, R.) 312–321 (PMLR, 2019).
26. Dgani, Y., Greenspan, H. & Goldberger, J. Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging* 39–42 (ISBI, 2018).
27. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med. Image Anal.* **65**, 101759 (2020).
28. Hendrycks, D., Mazeika, M., Wilson, D. & Gimpel, K. Using trusted data to train deep networks on labels corrupted by severe noise. In *Advances in Neural Information Processing Systems*, Vol. 31 (eds Bengio, S. et al.) (Curran Associates, Inc., 2018).
29. Krause, J. et al. The unreasonable effectiveness of noisy data for fine-grained recognition. In *Computer Vision – ECCV 2016* (eds Leibe, B., Matas, J., Sebe, N. & Welling, M.) 301–320 (Springer International Publishing, Cham, 2016).
30. Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 845–869 (2014).
31. Veit, A. et al. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR, 2017).
32. Yao, Q., Yang, H., Han, B., Niu, G. & Kwok, J. T.-Y. Searching to exploit memorization effect in learning with noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research* (eds III, H. D. & Singh, A.) 10789–10798 (PMLR, 2020).
33. Feldman, V. *Does Learning Require Memorization? A Short Tale about a Long Tail*. 954–959 (Association for Computing Machinery, New York, NY, USA, 2020).
34. Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S. & Vincent, P. The difficulty of training deep architectures and the effect of unsupervised pre-training. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Vol. 5 of *Proceedings of Machine Learning Research* (eds van Dyk, D. & Welling, M.) 153–160 (PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009).
35. Groschner, C. et al. Classifying handedness in chiral nanomaterials using label noise-robust deep learning. https://doi.org/10.5281/zenodo.5009042 (2021).

ACKNOWLEDGEMENTS

Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the US Department of Energy under Contract No. DE-AC02-05CH11231. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1752814. This work was also supported by National Science Foundation STROBE grant DMR-1548924. The authors wish to thank Alexander Mueller for collecting the SEM data.

AUTHOR CONTRIBUTIONS

M.C.S. and W.T. conceived the overall project. A.B.-M. synthesized the nanoparticles. C.K.G. implemented the mirror error labeling procedure and trained the neural networks. C.K.G. and A.J.P. optimized the networks. M.C.S. and C.K.G. wrote the manuscript. All authors commented on the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00822-7.

Correspondence and requests for materials should be addressed to M. C. Scott.

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.