Warwick Electron Microscopy Datasets

Jeffrey M. Ede¹, *

¹University of Warwick, Department of Physics, Coventry, CV4 7AL, UK
*j.m.ede@warwick.ac.uk

ABSTRACT

Large, carefully partitioned datasets are essential to train neural networks and standardize performance benchmarks. As a result, we have set up a new dataserver to make University of Warwick electron microscopy datasets available to the wider community. There are three main datasets containing 19769 scanning transmission electron micrographs, 17266 transmission electron micrographs, and 98340 simulated exit wavefunctions, with multiple variants of each dataset for different applications. Each dataset is visualized by t-distributed stochastic neighbour embedding, and we have created interactive visualization tools. Our datasets are supplemented by source code for analysis, data collection and visualization.

Datasets: https://warwick.ac.uk/fac/sci/physics/research/condensedmatt/microscopy/research/machinelearning
GitHub repository: https://github.com/Jeffrey-Ede/datasets

1 Introduction

We have set up a new dataserver to make our large new electron microscopy datasets available to both electron microscopists and the wider community. There are three main datasets containing 19769 experimental scanning transmission electron microscopy (STEM) images, 17266 experimental transmission electron microscopy (TEM) images and 98340 simulated TEM exit wavefunctions. Experimental datasets represent general research and were collected by dozens of University of Warwick scientists working on hundreds of projects between January 2010 and June 2018. We have been using our datasets to train artificial neural networks (ANNs) for electron microscopy, where standardizing results with common test sets has been essential for comparison. This paper provides details of and visualizations for datasets and their variants, and is supplemented by source code for analysis, data collection, and both static and interactive visualizations.

Machine learning is increasingly being applied to materials science, including to electron microscopy. Encouraging scientists, ANNs are universal approximators that can leverage an understanding of physics to represent the best way to perform a task with arbitrary accuracy. In theory, this means that ANNs can always match or surpass the performance of contemporary methods. However, training, validating and testing requires large, carefully partitioned datasets to ensure that ANNs are robust to general use. To this end, our datasets are partitioned so that each subset has different characteristics. For examples, by partitioning TEM or STEM images so that subsets are collected by different scientists, and by simulating exit wavefunction subsets with Crystallography Information Files (CIFs) for materials published in different journals.

Most areas of science are facing a reproducibility crisis, including artificial intelligence. Adding to this crisis, natural scientists do not always benchmark ANNs against standardized public domain test sets; making results difficult or impossible to compare. In electron microscopy, we believe this is a symptom of most datasets being small, esoteric or not having default partitions for machine learning. For example, most datasets in the Electron Microscopy Public Image Archive are for specific materials and are not partitioned. In contrast, standard machine learning datasets such as CIFAR-10, MNIST, and ImageNet have default partitions for machine learning and contain tens of thousands or millions of examples. By publishing our large, carefully partitioned machine learning datasets, and setting an example by using them to standardize our research, we aim to encourage higher standardization of machine learning research in the electron microscopy community.

There are many popular algorithms for high-dimensional data visualization. We use the scikit-learn implementation of t-distributed stochastic neighbour embedding (tSNE) as it is popular in the machine learning community. To reduce tSNE computation and data noise, we first apply probabilistic principal component analysis (PCA) to reduce the number of features in each image. This approach is used in the tSNE paper and works well in practice. Minka’s algorithm could be used to obtain the optimal number of principal components; however, that would require increased computation for singular value decomposition. As recommended by Oskolkov, we use tSNE perplexities given by $N^{1/2}$, where $N$ is the number of examples in a dataset, and confirm that changing perplexities by ±100 has little effect on visualizations for our large TEM and STEM datasets. Dataset details and visualizations are presented in the remaining sections of this paper.
Table 1. Examples and descriptions of STEM images in our datasets. References put some images into context to make them more tangible to unfamiliar readers.

2 Scanning Transmission Electron Micrographs

We curated 19769 STEM images from University of Warwick electron microscopy dataservers to train ANNs for compressed sensing\textsuperscript{5,7}. Atom columns are visible in roughly two-thirds of images, and similar proportions are bright and dark field. In addition, most signals are noisy\textsuperscript{48} and are imaged at several times their Nyquist rates\textsuperscript{49}. To reduce data transfer times for large images, we also created variant containing 161069 non-overlapping 512×512 crops from full images. For rapid development, we have also created new variants containing 96×96 images downsampled or cropped from full images. In this section we give details of each STEM dataset, referring to them using their names on our dataserver.

**STEM Full Images:** 19769 32-bit TIFFs containing STEM images taken with a University of Warwick JEOL ARM 200F electron microscope by dozens of scientists working on hundreds of projects. Images have their original sizes and intensities. Data was originally saved in DigitalMicrograph DM3 or DM4 files created by Gatan Microscopy Suite\textsuperscript{50} software, with tags containing rich metadata. However, metadata tags and original filenames have been removed from the public dataset to anonymize contributors. The dataset was made by concatenating contributions from different scientists, so partitioning the dataset before shuffling also partitions scientists.

**STEM Crops:** 161069 32-bit TIFFs containing 512×512 non-overlapping regions cropped from STEM Full Images. The dataset is partitioned into 110933 training, 21259 validation, and 28877 test set images. This dataset is biased insofar that larger images were divided into more crops.

**STEM 96×96:** A 32-bit NumPy\textsuperscript{51,52} array with shape [19769, 96, 96, 1] containing 19769 STEM Full Images area downsam-
pled to 96×96 with MATLAB and default antialiasing.

**STEM 96×96 Crops:** A 32-bit NumPy array with shape [19769, 96, 96, 1] containing 19769 96×96 regions cropped from STEM Full Images. Each crop is from a different image.

The distribution of STEM images is shown in figure 1 for STEM images downsampled to 96×96, and the distribution of structure in 96×96 crops from STEM images is shown in figure 2. Both visualizations were creating by embedding the first 50 principal components of images in two dimensions with tSNE. We used a perplexity of 127.4, 10000 iterations, and scikit-learn defaults for other parameters. Interactive visualizations that display images when you hover over map points are also available. This paper is aimed at a general audience so readers may not be familiar with STEM. Subsequently, example images are tabulated with references and descriptions in table 1 to make them more tangible.

### 3 Transmission Electron Micrographs

We curated 17266 2048×2048 TEM images from University of Warwick electron microscopy dataservers to train neural networks to improve signal-to-noise. However, our dataset was only available upon request. It is now available on our new dataserver. For convenience, we have also created a new variant containing 96×96 images that can be used for rapid ANN development. In this section we give details of each TEM dataset, referring to them using their names on our dataserver.

**TEM Full Images:** 17266 32-bit TIFFs containing 2048×2048 TEM images taken with University of Warwick JEOL 2000, JEOL 2100, JEOL 2100+, and JEOL ARM 200F electron microscope by dozens of scientists working on hundreds

| Lacey Carbon Supports | Apertures Blocking Electrons |
|-----------------------|-----------------------------|
| ![Lacey Carbon Supports](image1) | ![Apertures Blocking Electrons](image2) |

| Block Copolymers | Diffraction Patterns |
|------------------|----------------------|
| ![Block Copolymers](image3) | ![Diffraction Patterns](image4) |

| Vacuum at Specimen Edges | Nanowires |
|--------------------------|-----------|
| ![Vacuum at Specimen Edges](image5) | ![Nanowires](image6) |

| Multilayer Materials | Particles |
|----------------------|-----------|
| ![Multilayer Materials](image7) | ![Particles](image8) |

**Table 2.** Examples and descriptions of TEM images in our datasets. References put some images into context to make them more tangible to unfamiliar readers.
We simulated 98,340 TEM exit wavefunctions to train ANNs to reconstruct phases from amplitudes. Half of the wavefunctions were simulated at 512 × 512 then centre-cropped to 320 × 320 to remove simulation edge artefacts. Wavefunctions have been simulated for real physics where Kirkland potentials 29 for each atom are summed from \( n = 3 \) terms, and by truncating Kirkland potential summations to \( n = 1 \) to simulate an alternate universe where atoms have different potentials. Wavefunctions simulated for an alternate universe can be used to test ANN robustness to simulation physics. For rapid development, we also simulated for real physics where Kirkland potentials and a large range of materials and physical hyperparameters. The dataset is partitioned into 38,611 training, 3,861 validation, and 964 validation set wavefunctions. Metadata JSONs link wavefunctions to CIFs and contain some simulation hyperparameters.

To be clear, the training subset is at Python indexes [:24,530].

To show the distribution of TEM images in figure 3, we embedded the first 50 principal components of 96 × 96 images in two dimensions with tSNE. We used a perplexity of 131.4, 10,000 iterations, and scikit-learn defaults for other parameters. An interactive visualization that displays images when you hover over map points is also available 3. This paper is aimed at a general audience so readers may not be familiar with TEM. Subsequently, example images are tabulated with references and descriptions in table 2 to make them more tangible.

4 Exit Wavefunctions

We simulated 98,340 TEM exit wavefunctions to train ANNs to reconstruct phases from amplitudes 3. Half of wavefunction information is undetected by conventional TEM as only the amplitude, and not the phase, of an image is recorded. Wavefunctions were simulated at 512 × 512 then centre-cropped to 320 × 320 to remove simulation edge artefacts. Wavefunctions have been simulated for real physics where Kirkland potentials 29 for each atom are summed from \( n = 3 \) terms, and by truncating Kirkland potential summations to \( n = 1 \) to simulate an alternate universe where atoms have different potentials. Wavefunctions simulated for an alternate universe can be used to test ANN robustness to simulation physics. For rapid development, we also simulated for real physics where Kirkland potentials and a large range of materials and physical hyperparameters. The dataset is partitioned into 38,611 training, 3,861 validation, and 964 validation set wavefunctions. Metadata JSONs link wavefunctions to CIFs and contain some simulation hyperparameters.

To be clear, the training subset is at Python indexes [:24,530].

To show the distribution of TEM images in figure 3, we embedded the first 50 principal components of 96 × 96 images in two dimensions with tSNE. We used a perplexity of 131.4, 10,000 iterations, and scikit-learn defaults for other parameters. An interactive visualization that displays images when you hover over map points is also available 3. This paper is aimed at a general audience so readers may not be familiar with TEM. Subsequently, example images are tabulated with references and descriptions in table 2 to make them more tangible.

4 Exit Wavefunctions

We simulated 98,340 TEM exit wavefunctions to train ANNs to reconstruct phases from amplitudes 3. Half of wavefunction information is undetected by conventional TEM as only the amplitude, and not the phase, of an image is recorded. Wavefunctions were simulated at 512 × 512 then centre-cropped to 320 × 320 to remove simulation edge artefacts. Wavefunctions have been simulated for real physics where Kirkland potentials 29 for each atom are summed from \( n = 3 \) terms, and by truncating Kirkland potential summations to \( n = 1 \) to simulate an alternate universe where atoms have different potentials. Wavefunctions simulated for an alternate universe can be used to test ANN robustness to simulation physics. For rapid development, we also simulated for real physics where Kirkland potentials and a large range of materials and physical hyperparameters. The dataset is partitioned into 38,611 training, 3,861 validation, and 964 validation set wavefunctions. Metadata JSONs link wavefunctions to CIFs and contain some simulation hyperparameters.

To be clear, the training subset is at Python indexes [:24,530].

To show the distribution of TEM images in figure 3, we embedded the first 50 principal components of 96 × 96 images in two dimensions with tSNE. We used a perplexity of 131.4, 10,000 iterations, and scikit-learn defaults for other parameters. An interactive visualization that displays images when you hover over map points is also available 3. This paper is aimed at a general audience so readers may not be familiar with TEM. Subsequently, example images are tabulated with references and descriptions in table 2 to make them more tangible.

4 Exit Wavefunctions

We simulated 98,340 TEM exit wavefunctions to train ANNs to reconstruct phases from amplitudes 3. Half of wavefunction information is undetected by conventional TEM as only the amplitude, and not the phase, of an image is recorded. Wavefunctions were simulated at 512 × 512 then centre-cropped to 320 × 320 to remove simulation edge artefacts. Wavefunctions have been simulated for real physics where Kirkland potentials 29 for each atom are summed from \( n = 3 \) terms, and by truncating Kirkland potential summations to \( n = 1 \) to simulate an alternate universe where atoms have different potentials. Wavefunctions simulated for an alternate universe can be used to test ANN robustness to simulation physics. For rapid development, we also simulated for real physics where Kirkland potentials and a large range of materials and physical hyperparameters. The dataset is partitioned into 38,611 training, 3,861 validation, and 964 validation set wavefunctions. Metadata JSONs link wavefunctions to CIFs and contain some simulation hyperparameters.
components are at index [....,1]. The dataset can be partitioned in 24530 training, 3399 validation, and 8395 test set wavefunctions, which have been concatenated in that order. To be clear, the training subset is at Python indexes [:24530].

**wavefunctions**

- **restricted**
  - `n=3`: A 32-bit NumPy array with shape [11870, 96, 96, 2] containing 11870 wavefunctions. The wavefunctions were simulated for a large range of materials and a small range of physical hyperparameters, and bilinearly downsampled with skimage from 320×320 to 96×96 using default antialiasing. The dataset can be partitioned in 8002 training, 1105 validation, and 2763 test set wavefunctions, which have been concatenated in that order.

- **single**
  - `n=3`: A 32-bit NumPy array with shape [4825, 96, 96, 2] containing 11870 wavefunctions. The wavefunctions were simulated for In₁.₇K₂Se₈Sn₂.₂₈ and a large range of physical hyperparameters, and bilinearly downsampled with skimage from 320×320 to 96×96 using default antialiasing. The dataset can be partitioned in 3861 training, and 964 validation set wavefunctions, which have been concatenated in that order.

**unseen**

- **train**: 1501 64-bit NumPy files containing 320×320 wavefunctions. The wavefunctions are for a large range of materials and physical hyperparameters. The dataset is partitioned into 25352 training, 3569 validation, and 8563 test set wavefunctions. These wavefunctions are for an alternate universe where atoms have different potentials.

- **single**: 4819 complex 64-bit NumPy files containing 320×320 wavefunctions. The wavefunctions are for a single material, In₁.₇K₂Se₈Sn₂.₂₈, and a large range of physical hyperparameters. The dataset is partitioned into 3856 training, and 963 validation set wavefunctions. Metadata JSONs link wavefunctions to CIFs and contain some simulation hyperparameters. These wavefunctions are for an alternate universe where atoms have different potentials.

- **focal**: 1000 experimental focal series. Each series consists of 14 32-bit 512×512 TEM images, area downsampling from 4096×4096 with MATLAB and default antialiasing. The images are in TIFF format. All series were created with a common, quadratically increasing defocus series. However, spatial scales vary and must be fitted as part of reconstruction.

In detail, exit wavefunctions for a large range of physical hyperparameters were simulated with clTEM for acceleration voltages in {80, 200, 300} kV, material depths uniformly distributed in [5, 100] nm, material widths in [5, 10] nm, and crystallographic zone axes \((h,k,l)\) \(h,k,l \in \{0,1,2\}\). Materials were padded on all sides with vacuum 0.8 nm wide and 0.3 nm deep to reduce simulation artefacts. Finally, crystal tilts were perturbed by zero-centred Gaussian random variates with standard deviation 0.1°. We used default values for other clTEM hyperparameters. Simulations for a small range of physical hyperparameters used lower upper bounds that reduced simulation hyperparameter ranges by factors close to 1/4. All wavefunctions are linearly transformed to have a mean amplitude of 1.

Visualization of complex exit wavefunctions is complicated by the display of their real and imaginary components. However, real and imaginary components are related and can be visualized in the same image by plotting them in red and blue colour channels, respectively. Distributions of 96×96 simulated wavefunctions are shown in figure 4, figure 5, and figure 6 for a large range of materials and physical hyperparameters, a large range of materials and a small range of physical hyperparameters, and In₁.₇K₂Se₈Sn₂.₂₈ and a large range of physical hyperparameters, respectively. Visualizations were creating by embedding the first 50 principal components of amplitudes in two dimensions with tSNE. We used perplexities of 190.6, 108.9 and 69.5, respectively, 10000 iterations, and scikit-learn defaults for other parameters. Interactive visualizations that display amplitudes when you hover over map points are also available. All amplitude images show atom columns.

### 5 Discussion

The best dataset variant varies for different applications. Full-sized datasets can always be used as other dataset variants are derived from them. However, loading and processing full-sized examples may bottleneck training, and it is often unnecessary. Instead, smaller 512×512 crops, which can be loaded more quickly the full-sized images, can often be used to train ANNs to be applied convolutionally to or tiled across full-sized inputs. In addition, 96×96 dataset variants can be used in the early stages of development to rapidly train small ANNs. However, subtle application- and dataset-specific considerations may also influence the best dataset choice.

In practice, electron microscopists image most STEM and TEM signals at several times their Nyquist rates. This eases visual inspection, decreases sub-Nyquist aliasing, improves display on computer monitors, and is easier than carefully tuning...
sampling rates to capture the minimum data needed to resolve signals. High sampling may also reveal additional high-frequency information when images are inspected after an experiment. However, this complicates ANN development as it means that information per pixel is often higher in downsampled images. For example, partial scans across STEM images that have been dowsampled to $96 \times 96$ require higher coverages than scans across $96 \times 96$ crops for ANNs to learn to complete images with equal performance. It also complicates the comparison of different approaches to compressed sensing. For example, we suggested that sampling $512 \times 512$ crops at a regular grid of probing locations outperforms sampling along spiral paths as a subsampling grid can still access most information.

Test set performance should be calculated for a standardized dataset partition to ease comparison with other methods. Nevertheless, training and validation partitions can be varied to investigate validation variance for partitions with different characteristics. Default training and validation sets for STEM and TEM datasets contain contributions from different scientists that have been concatenated or numbered in order, so new validation partitions can be selected by concatenating training and validation partitions and moving the window used to select the validation set. Similarly, exit wavefunctions were simulated with CIFs from different journals that were concatenated or numbered sequentially. There is leakage between training, validation and test sets due to overlap between materials published in different journals and between different scientists’ work. However, further leakage can be minimized by selecting dataset partitions before any shuffling and, for wavefunctions, by ensuring that wavefunctions simulated for each journal are not split between partitions.

Experimental STEM and TEM image quality is variable. Images were taken by scientists with all levels of experience and TEM images were taken on multiple microscopes. This means that our datasets contain images that might be omitted from other datasets. For example, the tSNE visualization for STEM in figure 3 revealed some images where scans are incomplete, $\sim 50$ blank images, and images that only contain noise. Similarly, the tSNE visualization for TEM in figure 3 revealed some images where apertures block electrons, and that there are small number of unprocessed standard diffraction and convergent beam electron diffraction patterns. Although these conventionally low-quality images would not normally be published, they are important to ensure that ANNs are robust for live applications. We encourage readers to try our interactive tSNE visualizations for detailed inspection of our datasets.

Features extracted with PCA result in tSNE visualizations where image positions are related to total intensities. For example, in figure 1 dark and bright field STEM images are separated, and clusters at the top and bottom contain images with inverted intensity structure. Other approaches to image feature extraction could base visualizations on different image characteristics. For example, image features could be extracted with deep learning by using the latent space of an autoencoder or variational autoencoder, or by using features before logits in a classification ANN. We have posted autoencoders for electron microscopy with pre-trained models that could be improved. Alternatively, image features could be extracted from a histogram of oriented gradients, with speeded-up robust features, from local binary patterns, by wavelet decomposition or with other methods. The best features to extract for a visualization depend on its purpose.

6 Conclusion

We have provided details and visualizations for large new electron microscopy datasets available on our new dataserver. Datasets have been carefully partitioned into training, validation, and test sets for machine learning. In addition to full-sized datasets, we have provided variants containing $512 \times 512$ crops to reduce data loading times, and examples downsampled to $96 \times 96$ for rapid development. Source code and interactive dataset visualizations are provided in a supplementary repository to help users become familiar with our datasets. By making our datasets available, we aim to encourage standardization of performance benchmarks in electron microscopy and increase participation of the wider computer science community in electron microscopy research.

7 Data Availability

The data that support the findings of this study are openly available. Large new TEM, STEM, and exit wavefunctions datasets are on our new dataserver. Source code for data collection, figure preparation, and interactive visualizations are in a GitHub repository. For additional information contact the corresponding author (J.M.E.).

We do not have plans to host additional datasets from external users. However, we are open to hosting and encourage inquiry. We have funding for our public dataserver until at least 2030 and expect data to be moved to another archive if ours is no longer maintained. Datasets are accessed via hyperlinks on a main page so that we can change the physical locations of data without affecting users.

References

1. Ede, J. M. & Beanland, R. Electron Microscopy Datasets. Online: https://warwick.ac.uk/fac/sci/physics/research/condensedmatt/microscopy/research/machinelearning (2020).
2. FEI Company. An Introduction to Electron Microscopy. Online: https://www.fei.com/documents/introduction-to-microscopy-document (2010).

3. Ede, J. M., Peters, J. J. P., Sloan, J. & Beanland, R. Exit Wavefunction Reconstruction from Single Transmission Electron Micrographs with Deep Learning. *arXiv preprint arXiv:2001.10938* (2020).

4. Ede, J. M. & Beanland, R. Improving Electron Micrograph Signal-to-Noise with an Atrous Convolutional Encoder-Decoder. *Ultramicroscopy* **202**, 18–25 (2019).

5. Ede, J. M. & Beanland, R. Partial Scanning Transmission Electron Microscopy with Deep Learning. *arXiv preprint arXiv:1910.10467* (2020).

6. Ede, J. M. & Beanland, R. Adaptive Learning Rate Clipping Stabilizes Learning. *arXiv preprint arXiv:1906.09060* (2019).

7. Ede, J. M. Deep Learning Supersampled Scanning Transmission Electron Microscopy. *arXiv preprint arXiv:1910.10467* (2019).

8. Peters, J. J. P. & Dyson, M. A. Dataset Preparation and Visualization. Online: https://github.com/Jeffrey-Ede/datasets (2020).

9. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **5**, 1–36 (2019).

10. von Lilienfeld, O. A. Introducing Machine Learning: Science and Technology. *Mach. Learn. Sci. Technol.* **1**, 010201 (2020).

11. Belianinov, A. *et al.* Big Data and Deep Data in Scanning and Electron Microscopies: Deriving Functionality from Multidimensional Data Sets. *Adv. Struct. Chem. Imaging* **1**, 1–25 (2015).

12. Hornik, K., Stinchcombe, M. & White, H. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks* **2**, 359–366 (1989).

13. Lin, H. W., Tegmark, M. & Rolnick, D. Why does Deep and Cheap Learning Work so Well? *J. Stat. Phys.* **168**, 1223–1247 (2017).

14. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv preprint arXiv:1811.12808* (2018).

15. Roh, Y., Heo, G. & Whang, S. E. A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective. *IEEE Transactions on Knowl. Data Eng.* (2019).

16. Hall, S. R., Allen, F. H. & Brown, I. D. The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography. *Acta Crystallogr. Sect. A: Foundations Crystallogr.* **47**, 655–685 (1991).

17. Baker, M. Reproducibility Crisis? *Nature* **533**, 353–66 (2016).

18. Hutson, M. Artificial Intelligence Faces Reproducibility Crisis (2018).

19. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: A Public Archive for Raw Electron Microscopy Image Data. *Nat. Methods* **13**, 387 (2016).

20. Hey, T., Butler, K., Jackson, S. & Thiyagalingam, J. Machine Learning and Big Scientific Data. *Philos. Transactions Royal Soc. A* **378**, 20190054 (2020).

21. Krizhevsky, A., Nair, V. & Hinton, G. The CIFAR-10 Dataset. Online: http://www.cs.toronto.edu/~kriz/cifar.html (2014).

22. Krizhevsky, A. & Hinton, G. Learning Multiple Layers of Features from Tiny Images. Tech. Rep., Citeseer (2009).

23. LeCun, Y., Cortes, C. & Burges, C. MNIST Handwritten Digit Database. AT&T Labs, online: http://yann.lecun.com/exdb/mnist (2010).

24. Russakovsky, O. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).

25. Tenenbaum, J. B., De Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290**, 2319–2323 (2000).

26. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *science* **290**, 2323–2326 (2000).

27. Zhang, Z. & Wang, J. MLLE: Modified Locally Linear Embedding Using Multiple Weights. In *Advances in Neural Information Processing Systems*, 1593–1600 (2007).
28. Donoho, D. L. & Grimes, C. Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data. *Proc. Natl. Acad. Sci.* **100**, 5591–5596 (2003).
29. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **15**, 1373–1396 (2003).
30. Zhang, Z. & Zha, H. Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *SIAM J. on Sci. Comput.* **26**, 313–338 (2004).
31. Buja, A. *et al.* Data Visualization with Multidimensional Scaling. *J. Comput. Graph. Stat.* **17**, 444–472 (2008).
32. Van Der Maaten, L. Accelerating t-SNE Using Tree-Based Algorithms. *The J. Mach. Learn. Res.* **15**, 2825–2830 (2011).
33. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
34. Halko, N., Martinsson, P.-G. & Tygert, M. A Randomized Algorithm for the Decomposition of Matrices. *Appl. Comput. Harmon. Analysis* **30**, 47–68 (2011).
35. Oskolkov, N. How to Tune Hyperparameters of tSNE. *Towards Data Science*, online: https://towardsdatascience.com/how-to-tune-hyperparameters-of-tsne-7c0596a18868 (2019).
36. Seki, T., Ikuhara, Y. & Shibata, N. Theoretical Framework of Statistical Noise in Scanning Transmission Electron Microscopy. *Ultramicroscopy* **193**, 118–125 (2018).
37. Karlsson, G. Thickness Measurements of Lacey Carbon Films. *J. Microsc.* **203**, 326–328 (2001).
38. Inam, M. *et al.* 1D vs. 2D Shape Selectivity in the Crystallization-Driven Self-Assembly of Polylactide Block Copolymers. *Chem. Sci.* **8**, 4223–4230 (2017).
39. Bendersky, L. A. & Gayle, F. W. Electron Diffraction Using Transmission Electron Microscopy. *J. Res. Natl. Inst. Standards Technol.* **106**, 997 (2001).
56. Wu, Y., Messer, B. & Yang, P. Superconducting MgB$_2$ Nanowires. *Adv. Mater.* **13**, 1487–1489 (2001).
57. Pang, B. *et al*. The Microstructural Characterization of Multiferroic LaFeO$_3$-YMnO$_3$ Multilayers Grown on (001) and (111)-SrTiO$_3$ Substrates by Transmission Electron Microscopy. *Materials* **10**, 839 (2017).
58. Dong, Z. *et al*. Individual Particles of Cryoconite Deposited on the Mountain Glaciers of the Tibetan Plateau: Insights into Chemical Composition and Sources. *Atmospheric Environ.* **138**, 114–124 (2016).
59. Kirkland, E. J. *Advanced Computing in Electron Microscopy* (Springer Science & Business Media, 2010).
60. ISO/IEC JTC 1/SC 22. International Standard ISO/IEC21778: Information Technology - The JSON Data Interchange Syntax. Online: https://www.iso.org/standard/71616.html (2017).
61. Hwang, S.-J., Iyer, R. G., Trikalitis, P. N., Ogden, A. G. & Kanatzidis, M. G. Cooling of Melts: Kinetic Stabilization and Polymorphic Transitions in the KInSnSe$_4$ System. *Inorg. chemistry* **43**, 2237–2239 (2004).
62. Van der Walt, S. *et al*. scikit-image: Image Processing in Python. *PeerJ* **2**, e453 (2014).
63. Python Software Foundation. Python 3.6. Online: http://www.python.org (2020).
64. Adobe Developers Association *et al*. TIFF Revision 6.0. Online: www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf (1992).
65. Haigh, S., Jiang, B., Alloypeau, D., Kiesielowski, C. & Kirkland, A. Recording Low and High Spatial Frequencies in Exit Wave Reconstructions. *Ultramicroscopy* **133**, 26–34 (2013).
66. Peters, J. J. P. & Dyson, M. A. cITEM. Online: https://github.com/JJPPeters/cITEM (2019).
67. Dyson, M. A. *Advances in Computational Methods for Transmission Electron Microscopy Simulation and Image Processing*, Ph.D. thesis, University of Warwick (2014).
68. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232 (2017).
69. Amidror, I. Sub-Nyquist Artefacts and Sampling Moiré Effects. *Royal Soc. Open Sci.* **2**, 140550 (2015).
70. Open Data Science. How to Fix Data Leakage - Your Model’s Greatest Enemy. Towards Data Science, online: https://medium.com/@ODSC/how-to-fix-data-leakage-your-models-greatest-enemy-e34fa26abac5 (2019).
71. Bussola, N., Marcolini, A., Maggio, V., Jurman, G. & Furlanello, C. Not Again! Data Leakage in Digital Pathology. *arXiv preprint arXiv:1909.06539* (2019).
72. Tanaka, M. Convergent-Beam Electron Diffraction. *Acta Crystallogr. Sect. A: Foundations Crystallogr.* **50**, 261–286 (1994).
73. Tschannen, M., Bachem, O. & Lucic, M. Recent Advances in Autoencoder-Based Representation Learning. *arXiv preprint arXiv:1812.05069* (2018).
74. Kramer, M. A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE J.* **37**, 233–243 (1991).
75. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114* (2013).
76. Marcelino, P. Transfer Learning From Pre-trained Models. Towards Data Science, online: https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751 (2018).
77. Ede, J. M. Kernels, MLPs and Autoencoders. Online: https://github.com/Jeffrey-Ede/Denoising-Kernels-MLPs-Autoencoders (2018).
78. Ede, J. M. Autoencoders, Kernels, and Multilayer Perceptrons for Electron Micrograph Restoration and Compression. *arXiv preprint arXiv:1808.09916* (2018).
79. Dalal, N. & Triggs, B. Histograms of Oriented Gradients for Human Detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, 886–893 (IEEE, 2005).
80. Bay, H., Ess, A., Tuytelaars, T. & Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **110**, 346–359 (2008).
81. Ojala, T., Pietikainen, M. & Maenpaa, T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on pattern analysis machine intelligence* **24**, 971–987 (2002).
82. Mallat, S. G. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis Mach. Intell.* **11**, 674–693 (1989).
83. Latif, A. et al. Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review. *Math. Probl. Eng.* 2019 (2019).

8 Acknowledgements

Thanks go to Christoph T. Koch for software used to collect experimental focal series, to Richard Beanland for help setting up dataservers and to Chris Parkin for managing dataservers.

Funding: J.M.E. acknowledges EPSRC EP/N035437/1 for financial support. In addition, J.M.E. acknowledges EPSRC Studentship 1917382.

9 Competing Interests

The authors declare no competing interests.
Figure 1. Two-dimensional tSNE visualization of the first 50 principal components of 19769 STEM images that have been downsamped to 96×96. The same grid is used to show a) map points and b) images at 500 randomly selected points.
Figure 2. Two-dimensional tSNE visualization of the first 50 principal components of 19769 96×96 crops from STEM images. The same grid is used to show a) map points and b) images at 500 randomly selected points.
Figure 3. Two-dimensional tSNE visualization of the first 50 principal components of 17266 TEM images that have been downsampled to $96 \times 96$. The same grid is used to show a) map points and b) images at 500 randomly selected points.
Figure 4. Two-dimensional tSNE visualization of the first 50 principal components of 36324 exit wavefunctions that have been downsampled to $96 \times 96$. Wavefunctions were simulated for thousands of materials and a large range of physical hyperparameters. The same grid is used to show a) map points and b) wavefunctions at 500 randomly selected points. Red and blue colour channels show real and imaginary components, respectively.
Figure 5. Two-dimensional tSNE visualization of the first 50 principal components of 11870 exit wavefunctions that have been downsampled to $96 \times 96$. Wavefunctions were simulated for thousands of materials and a small range of physical hyperparameters. The same grid is used to show a) map points and b) wavefunctions at 500 randomly selected points. Red and blue colour channels show real and imaginary components, respectively.
Figure 6. Two-dimensional tSNE visualization of the first 50 principal components of 4825 exit wavefunctions that have been downsampled to $96 \times 96$. Wavefunctions were simulated for thousands of materials and a small range of physical hyperparameters. The same grid is used to show a) map points and b) wavefunctions at 500 randomly selected points. Red and blue colour channels show real and imaginary components, respectively.