Lawrence Berkeley National Laboratory
LBL Publications

Title
Contrast Transfer Functions Help Quantify Neural Network Out-Of-Distribution Generalization in HRTEM

Permalink
https://escholarship.org/uc/item/2n05f9z5

Journal
Microscopy and Microanalysis, 32(4)

ISSN
1431-9276

Authors
DaCosta, Luis Rangel
Scott, Mary C

Publication Date
2026-07-01

DOI
10.1093/mam/ozag067

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Contrast Transfer Functions Help Quantify Neural Network Out-Of-Distribution Generalization in HRTEM

Luis Rangel DaCosta, Mary C Scott

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

DECTRIS QUADRO

The flexible standard for **electron microscopy**
Maximizing performance on a single platform



Microscopy and Microanalysis, 2026, **32**, ozag067
https://doi.org/10.1093/mam/ozag067
Advance access publication 15 July 2026
Original Article

Microscopy AND Microanalysis

Contrast Transfer Functions Help Quantify Neural Network Out-Of-Distribution Generalization in HRTEM

Luis Rangel DaCosta[1,2,*], **and Mary C. Scott**[1,2,*]

[1]*Department of Materials Science and Engineering, University of California, Hearst Memorial Mining Building, Berkeley, CA 94720, USA*
[2]*National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720, USA*

***Corresponding authors:** Luis Rangel DaCosta, E-mails: lrangeldacosta@lbl.gov; Mary C. Scott, mary.scott@berkeley.edu

Abstract

Neural networks, while effective for tackling many challenging scientific tasks, are not known to perform well out-of-distribution (OOD), i. within domains which differ from their training data. Understanding neural network OOD generalization is paramount to their successful deployment in experimental workflows, especially when ground-truth knowledge about the experiment is hard to establish or experimental conditions significantly vary. With inherent access to ground-truth information and fine-grained control of underlying distributions, simulation-based data curation facilitates precise investigation of OOD generalization behavior. Here, we probe generalization with respect to imaging conditions of neural network segmentation models for high-resolution transmission electron microscopy (HRTEM) imaging of nanoparticles, training and measuring the OOD generalization of over 12,000 neural networks using synthetic data generated via random structure sampling and multislice simulation. Using the HRTEM contrast transfer function, we further develop a framework to compare information content of HRTEM datasets and quantify OOD domain shifts. We demonstrate that neural network segmentation models enjoy significant performance stability, but will smoothly and predictably worsen as imaging conditions shift from the training distribution. Lastly, we consider limitations of our approach in explaining other OOD shifts, such as of the atomic structures, and discuss complementary techniques for understanding generalization in such settings.

Key words: generalization, HRTEM, neural networks, segmentation

Introduction

The out-of-distribution (OOD) generalization capability of neural networks underpins the utility of machine learning models in experimental settings. OOD generalization considers performance on data that differs from the training distribution and we can broadly consider machine learning models to generalize well OOD when they can be used throughout large data domains without meaningful performance degradation. Many key advances in machine learning have worked toward developing an understanding of in-distribution generalization—the ability of a neural network model to learn near-optimal functions for analyzing data similar to their training distribution—but our comparative understanding of OOD generalization lags severely behind (Liu et al., 2023; Yu et al., 2024), largely relying on empirical evidence and benchmark studies in restricted domains (Kang et al., 2024; Ye et al., 2022). Few tools currently exist which can predict and quantitatively describe under which scenarios and to what extent a trained neural network model will successfully generalize to new target domains. As neural network models become increasingly adopted within experimental microscopy workflows, it is becoming increasingly important to devote effort to understanding OOD generalization phenomena as it applies to characterization.

What do OOD datasets look like for electron microscopy? Here, we strictly categorize in-distribution and OOD data in a probabilistic setting: data are in distribution with respect to a training dataset if and only if samples of the data are drawn independently and in an identically distributed (i.i.d) manner. If the data generating distribution changes in any way—and thus, samples are no longer drawn i.i.d.—such data are OOD, even if the generating distributions differ only slightly. For example, in supervised learning, a pair of training and validation datasets may consist of unique atomic structures—i.e., no individual structure commonly appears in both—but which are randomly sampled via the same underlying process. For these datasets, the validation structures are in-distribution with respect to the training structures. On the other hand, if the validation dataset was acquired under different imaging conditions—say, with a applied different defocus or amount of dose—the images themselves are now OOD, even though the structures are not. Because OOD shifts can be negligibly small, encountering OOD datasets does not immediately indicate that a machine learning model will fail. Instead, we can understand OOD degradation to occur more continuously, and such our priority shifts toward understanding which factors contribute to negatively impactful OOD shifts, and to what extent. While it may be practically safer to broadly consider OOD experimental data unsafe to analyze with a pretrained model, this may be overly conservative, if it can be determined that a trained model will stably generalize under expected OOD shifts.

OOD generalization behavior can be explicitly evaluated in supervised learning settings where supervision labels can be used to directly compute model performance, given that suitable OOD benchmark data are accessible. Such benchmark data, however, are not readily available in many experimental

Received: December 10, 2025. **Revised:** May 14, 2026. **Accepted:** June 23, 2026
© The Author(s) 2026. Published by Oxford University Press on behalf of the Microscopy Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

settings. Experimental acquisition of domain-shifted benchmark data in transmission electron microscopy (TEM) is feasible along some more controllable aspects of our datasets, like imaging resolution or nanoparticle chemistry, but is expensive to obtain and annotate. Indeed, in previous work (Sytwu et al., 2024), we have successfully explored neural network segmentation performance OOD via experimental acquisition with respect to shifts in such aspects, including image magnification, electron dose, nanoparticle size, and nanoparticle crystal structure, which required extensive experimental control and manual analysis even for the limited domains explored. Acquiring and annotating benchmark data for more complex analysis tasks which might be challenging for expert practitioners, or for which aspects of domain shift could be difficult to describe, measure, or calibrate, quickly becomes infeasible.

Alternatively, simulation-based data generation approaches (Rangel DaCosta et al., 2024) provide potentially unlimited data with ground-truth annotations and offer an attractive alternative for studying machine learning model generalization. Electron microscopists have recently rapidly adopted neural networks for atomic resolution analysis workflows with the TEM (Madsen et al., 2018; Groschner et al., 2021; Sadre et al., 2021; Vincent et al., 2021; Wang et al., 2021; Ede, 2021; Munshi et al., 2022; Lin et al., 2021; Lee et al., 2025; McCray et al., 2025; Abebe et al., 2025; Kim et al., 2025; Lobato et al., 2024; Belardi et al., 2025; Khan et al., 2023). Many such prior studies emphasize and explore the need for robust, generalizable models. For core, benchmark tasks such as semantic segmentation for analysis of high-resolution (or, phase-contrast) transmission electron microscopy (HRTEM) data, this work has already been quite extensive (Madsen et al., 2018; Groschner et al., 2021; Sadre et al., 2021; Sytwu et al., 2022; Larsen et al., 2023; Sytwu et al., 2024; Rangel DaCosta et al., 2024; Borshon et al., 2025), yet remains largely empirical and limited in scope, seldom offering physics-based or otherwise causal explanations for understanding model performance in-distribution and OOD. Alongside well-established tools for image simulation (Madsen & Susi, 2021; Rangel DaCosta et al., 2021), there exists a unique opportunity in electron microscopy to more fully develop our understanding of the limitations of neural network models for experimental data analysis and their potential to generalize across experimental scenarios.

Information theory underlies many of our working models of datasets and learning theory in machine learning and forms the classical basis for understanding OOD generalization of neural network models as a phenomenon arising from domain shift (Ben-David et al., 2010). Recent approaches extending such information theoretic perspective with ideas from causal modeling—specifically, the invariance of learnable features (Arjovsky et al., 2020; Ahuja et al., 2021; Ye et al., 2021)—have shown promise in providing a unified perspective of OOD generalization across learning tasks. While there exists some purely model- or optimization-centric methods to detect OOD data or improve robustness to OOD shifts (Ramé et al., 2022; Rame et al., 2022), such approaches do not readily predict the extent of failure of a model OOD and often rely on expensive computations. In contrast, data-centric approaches utilizing probabilistic and information theoretic perspectives (Liu et al., 2022; Miller et al., 2021; Shi et al., 2024; Cohen-Wang et al., 2025) develop an understanding of OOD performance via inter-dataset relationships and their relation to model behavior, laying out a route toward a more direct, mechanistic understanding of the OOD behavior of neural networks used for TEM analysis.

Within electron microscopy, concepts in information theory most commonly arise when considering information—or contrast—transfer functions. Contrast transfer functions describe how information propagates from samples to measurements via the system of imaging lenses, forming the foundation of our theories of image formation and techniques for interpreting the information contained within micrographs (Rose, 1984; Van Dyck & de Jong, 1992; de Jong & Van Dyck, 1993; Nellist et al., 1995; Spence, 2017). Due to the confluence of consumer-available high-performance computing, ultrafast detectors, and advanced ptychographic reconstruction algorithms (Chen et al., 2021; Küçükoğlu et al., 2024; Varnavides et al., 2024) in recent years, understanding information transfer has been considered important for determining limits of the modern TEM (Varnavides et al., 2024, 2026; Ma et al., 2026). For TEM imaging data, the contrast transfer function provides a natural setting for understanding image datasets and their information content from a distributional perspective.

In this work, we take a detailed look at the OOD generalization behavior of neural network models trained via supervised learning for HRTEM analysis, focusing on semantic segmentation of atomic-resolution micrographs of nanoparticles on substrates. Specifically, we study the task of segmenting nanoparticle regions, i.e., classifying contiguous regions of micrographs as either "nanoparticle" or "substrate"—for an example segmentation of atomic resolution HRTEM data, see Figure 1.

We begin by seeking to understand neural network OOD behavior with respect to shifts in experimental imaging conditions, which are experimental factors we can describe analytically and control moderately well, but which are hard to precisely measure and can shift significantly between different experiments or TEM operators. Using fully simulated datasets, whose imaging conditions we can arbitrarily control, we demonstrate that atomic resolution segmentation networks have stable and predictable degradation of performance when evaluating OOD data. Moreover, the extent of performance degradation can be mechanistically understood from an information theoretic perspective, by relating HRTEM training and test datasets via heuristic metrics we design, which provide comparisons between datasets using the familiar contrast transfer function (CTF). We believe this to be of the first use of CTFs to attempt to explain and predict OOD generalization of neural network models for TEM applications. Within this CTF-comparison framework, we establish data-centric settings with under which model performance will appreciably degrade. Lastly, we discuss extensions and limitations of our approach for understanding the effect of distribution shifts across different experimental settings, especially when considering shifts in the distribution of atomic structures.

Methods

Information Transfer Metrics

We will look to compare the imaging conditions between two HRTEM datasets via their corresponding contrast transfer functions. To simplify the form of the contrast transfer functions, we approximate our sample as a phase object. Throughout, we will use $\mathbf{r}$ to represent real space coordinates and $\mathbf{q}$ to represent reciprocal space coordinates.

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

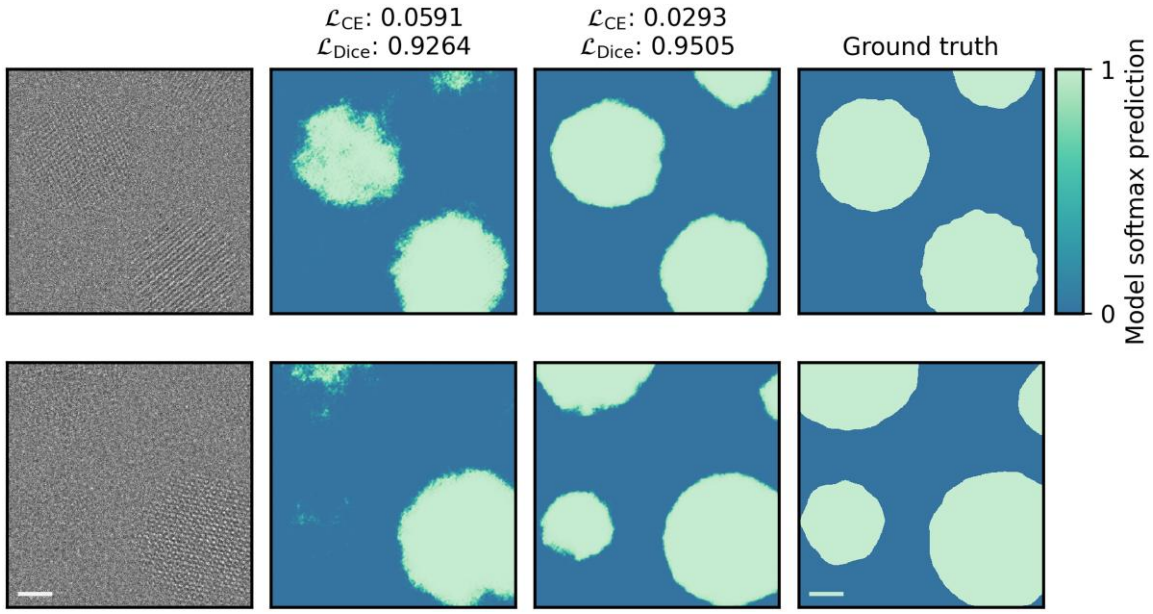


Fig. 1. Example pixel-wise semantic segmentation performance of a neural network at two different stages of optimization on images of CdSe nanoparticles with randomly sampled imaging conditions, with validation cross-entropy loss and Dice scores given for reference. Simulation-based ground truth, used for model supervision, is in the right most column. Scalebar is 1 nm.

After propagation through the sample and passing through the objective lens, the wavefunction at the back focal plane $\psi_f(\mathbf{q})$ for HRTEM imaging of phase objects can be calculated as

$$\psi_f(\mathbf{q}) = \underset{\mathbf{r}\to\mathbf{q}}{\mathcal{F}}\left[\psi_0(\mathbf{r})\exp\left(i\sigma V(\mathbf{r})\right)\right]H(\mathbf{q}), \tag{1}$$

where $\psi_0(\mathbf{r})$ is an incident plane wave, $H(\mathbf{q})$ is the transfer function, $V(\mathbf{r})$ is the electrostatic potential of the sample, and $\underset{\mathbf{r}\to\mathbf{q}}{\mathcal{F}}$ is the Fourier transform from real to reciprocal space. $H(\mathbf{q})$ directly modulates the wavefunction and is the product of several components (Williams & Carter, 2008):

$$H(\mathbf{q}) = A(\mathbf{q})E(\mathbf{q})\exp\left(-i\chi(\mathbf{q})\right). \tag{2}$$

$A(\mathbf{q})$ is an aperture function, which defines the maximum angular cutoff of incident electrons in the microscope due to a limiting aperture; $E(\mathbf{q})$ is a damping envelope function, which encompasses high-frequency damping arising from a variety of experimental sources; and $\exp\left(i\chi(\mathbf{q})\right)$ is known as the phase contrast transfer function. If we further assume that our sample is a weak phase object, we can instead utilize the intensity transfer function $T(\mathbf{q})$, which corresponds to measured image intensities $I(\mathbf{r}) = \psi_f(\mathbf{r})\overline{\psi_f(\mathbf{r})}$ and can be written as (Williams & Carter, 2008)

$$T(\mathbf{q}) = 2E(\mathbf{q})A(\mathbf{q})\sin\left(\chi(\mathbf{q})\right). \tag{3}$$

$T(\mathbf{q})$—also known as the objective lens transfer function—determines the ultimate resolution and information limits of the microscope, and can be thought as describing the extent to which information from the sample scattering from different angles contributes to the information contained in the final measured image. $T(\mathbf{q})$ decays at high frequencies and is oscillatory, which has traditionally complicated the interpretation of the contrast present in HRTEM micrographs. In spherical aberration-corrected microscopes, both ultimate resolution and information limits typically arise from high spatial frequency damping cause by chromatic aberration of the electron gun (Kirkland, 2020; Williams & Carter, 2008)

$$E(\mathbf{q}) = \exp\left(-(\pi\lambda\delta)^2\|\mathbf{q}\|^4/2\right), \tag{4}$$

where δ is the defocus spread due to chromatic aberration. The general form of $E(\mathbf{q})$ can be quite complex—e.g., by including effects of partial coherence (Spence, 2017)—but we will consider contributions to $E(\mathbf{q})$ arising only from chromatic aberration.

While resolution limits arise mostly due to the damping envelope $E(\mathbf{q})$, the qualitative form of $T(\mathbf{q})$ and thus the behavior of information transfer from sample to image is instead dominated by the aberrations of the objective lens. The objective lens aberrations are captured in the aberration phase shift $\chi(\mathbf{q})$, which, following (Rangel DaCosta et al., 2021), we define as

$$\begin{aligned}\chi(\mathbf{q}) = \sum_{m,n}(\lambda\|\mathbf{q}\|)^m C_{mn}^{\text{mag}}\big[&\cos\left(nC_{\theta,mn}\right)\cos\left(n\mathbf{q}_\theta\right)\\ &+\sin\left(nC_{mn}^{\theta}\right)\sin\left(n\mathbf{q}_\theta\right)\big].\end{aligned} \tag{5}$$

$\mathbf{q}_\theta$ is the polar angle of $\mathbf{q}$ as returned by the two-argument arctangent function, λ is the electron wavelength, and C_{mn}^{mag} and C_{mn}^{θ} are dimensionless coefficients corresponding to the aberration magnitude and azimuthal phase for an aberration of radial order m and azimuthal order n. Equation 5 can be further reduced with the angle-sum trigonometric identity to

$$\chi(\mathbf{q}) = \sum_{m,n}\left[(\lambda\|\mathbf{q}\|)^m C_{m,n}^{\text{mag}}\cos\left(n(C_{m,n}^{\text{ang}} - \mathbf{q}_\theta)\right)\right]. \tag{6}$$

We distinguish two sets of imaging conditions only by their lens aberrations and characterize their differences with respect to their corresponding information transfer properties via two auxiliary heuristic metrics, $\epsilon(\chi(\mathbf{q}))$ and $\sigma(\chi(\mathbf{q}), \chi'(\mathbf{q}))$, derived from $T(\mathbf{q})$. Since it is usually unambiguous, for brevity we

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

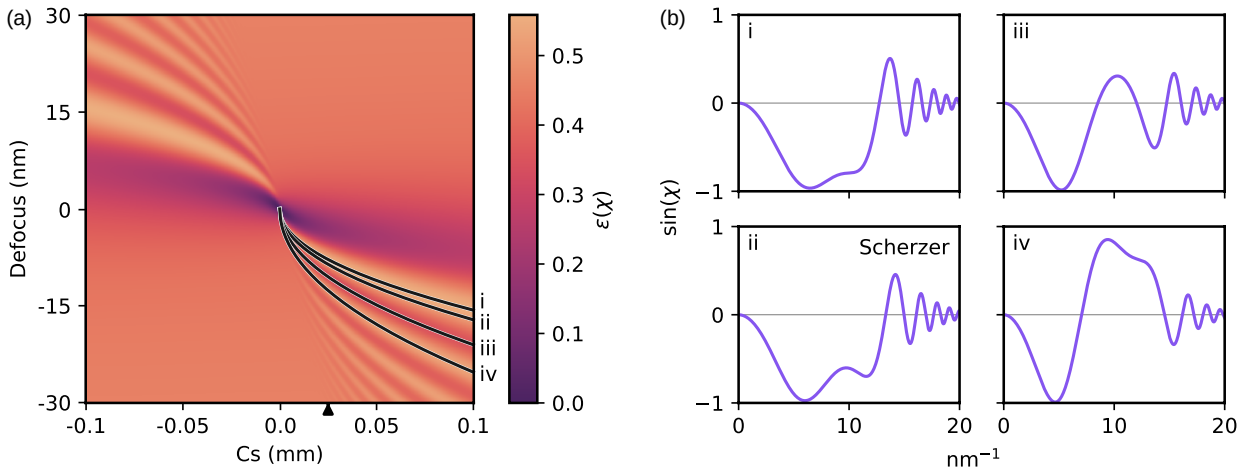


Fig. 2. (**a**) Example visualization of $\epsilon(\chi)$ for aberrations consisting solely of defocus and spherical aberrations, with magnitudes up to 30 nm and 0.1 mm, respectively, at 300 kV and with a chromatic aberration envelope corresponding to a focal spread of 10Å. Bright bands in regions where defocus and spherical aberration have opposite sign correspond to the formation of passbands in the contrast transfer function. (**b**) A selection of phase contrast transfer functions corresponding to maxima and minima in (**a**, i–iv) are visualized for a lens with a spherical aberration coefficient of 25 μm, indicated by the black triangle marker.

write $\epsilon(\chi) = \epsilon(\chi(\mathbf{q}))$ and $\sigma(\chi, \chi') = \sigma(\chi(\mathbf{q})), \chi'(\mathbf{q}))$. These metrics were designed to be simple heuristic measures for intuitively relating contrast transfer functions, themselves a familiar measure in HRTEM imaging, but we note that they do not necessarily correspond to explicit measures or comparisons of the information content within HRTEM images.

We define the first metric, $\epsilon(\chi)$, as

$$\epsilon(\chi) = \frac{\int |T(\mathbf{q})|^2 d\mathbf{q}}{\int |E(\mathbf{q})|^2 d\mathbf{q}} \tag{7}$$

which can be thought of as measuring the efficiency of information transfer from the sample to the image with respect to a perfect contrast transfer scenario when limited by high frequency damping (i.e., $T(\mathbf{q}) = 2E(\mathbf{q})$).

The second metric, $\sigma(\chi, \chi')$, is defined as

$$\sigma(\chi, \chi') = \frac{\int |T(\mathbf{q})||T'(\mathbf{q})| d\mathbf{q}}{\int |T(\mathbf{q})|^2 d\mathbf{q}}. \tag{8}$$

$\sigma(\chi, \chi')$ can be thought as an overlap metric which measures the similarity between two contrast transfer functions. From the perspective of an imaging dataset, $\sigma(\chi, \chi')$ helps describe the extent to which images acquired under a new imaging condition χ' (e.g., a test dataset) capture information from the sample in similar proportions across relevant spatial frequencies as the first, χ (e.g., a training dataset). We normalize $\sigma(\chi, \chi')$ by the maximum possible overlap (i.e., the training dataset with itself) to be able to asymmetrically compare imaging conditions, which will be helpful in understanding the necessarily asymmetric relationships between training and test datasets in neural network training.

Here, we use $|T(\mathbf{q})$ instead of $T(\mathbf{q})$ to consider only the amount of information transferred at each frequency, which has no effect on $\epsilon(\chi)$. However, this restricts $\sigma(\chi, \chi')$ to be non-negative, and further, to be invariant to contrast reversals. In this framing, for example, two images acquired with the same magnitude and opposite sign of defocus are equivalent, in terms of their information content and its dependence on spatial frequency (with respect to the sample), even though the images themselves may look quite different. For comparison, one could consider this to be analogous in interpretation to the difference between the (square root of the) power spectrum of a (real) signal and the signal's Fourier transform.

To demonstrate its use, in Figure 2a, we show how $\epsilon(\chi)$ varies across imaging conditions varying in defocus and spherical aberration. When imaging with both defocus and spherical aberration close to zero, $\epsilon(\chi)$ is also close to zero—no information is transferred from the sample in an aberration-free lens —and when defocus and spherical aberration have the same sign, $\epsilon(\chi)$ has almost constant value. It becomes possible to form passbands in the contrast transfer function (Spence, 2017) when spherical aberration and defocus have opposite sign, and in turn, these passbands form local maxima of $\epsilon(\chi)$. We note here that these local maxima do not correspond to the widest passbands—those with the largest windows of interpretable image frequencies, such as Scherzer defocus—but are rather slightly periodically offset from such the sequence of passbands (Fig. 2b).

In Figure 3, we further show how $\epsilon(\chi)$ and $\sigma(\chi, \chi')$ together can be used to compare imaging conditions, considering a simple case in which two HRTEM datasets vary only in defocus. Generally, $\sigma(\chi, \chi')$ is maximized when information is transferred proportionally at similar frequencies, i.e., when the contrast transfer functions of two datasets closely align, which occurs here when the magnitudes of defocus are similar between train and test datasets (Fig. 3a,c). We use $\epsilon(\chi)$ to compare the total information content between HRTEM two datasets via their difference $\Delta\epsilon(\chi, \chi') = \epsilon(\chi') - \epsilon(\chi)$ (Fig. 3b), which we note is not sensitive to the spatial frequency dependence of the two transfer functions. A particularly extreme case can occur when either the training or testing dataset comes from near an aberration-free condition—in Figure 3, these are conditions similar to that indicated by the triangle marker. When the training dataset nears zero defocus, the contrast transfer function decays in magnitude and loses its oscillations, causing $\sigma(\chi, \chi') \geq 1$ and further implying that the training dataset has strong overlap to nearly any other aberration condition. Conversely, $\sigma(\chi, \chi') \to 0$ when instead the test dataset nears zero defocus. Such conditions also cause $\Delta\epsilon(\chi, \chi')$ to

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

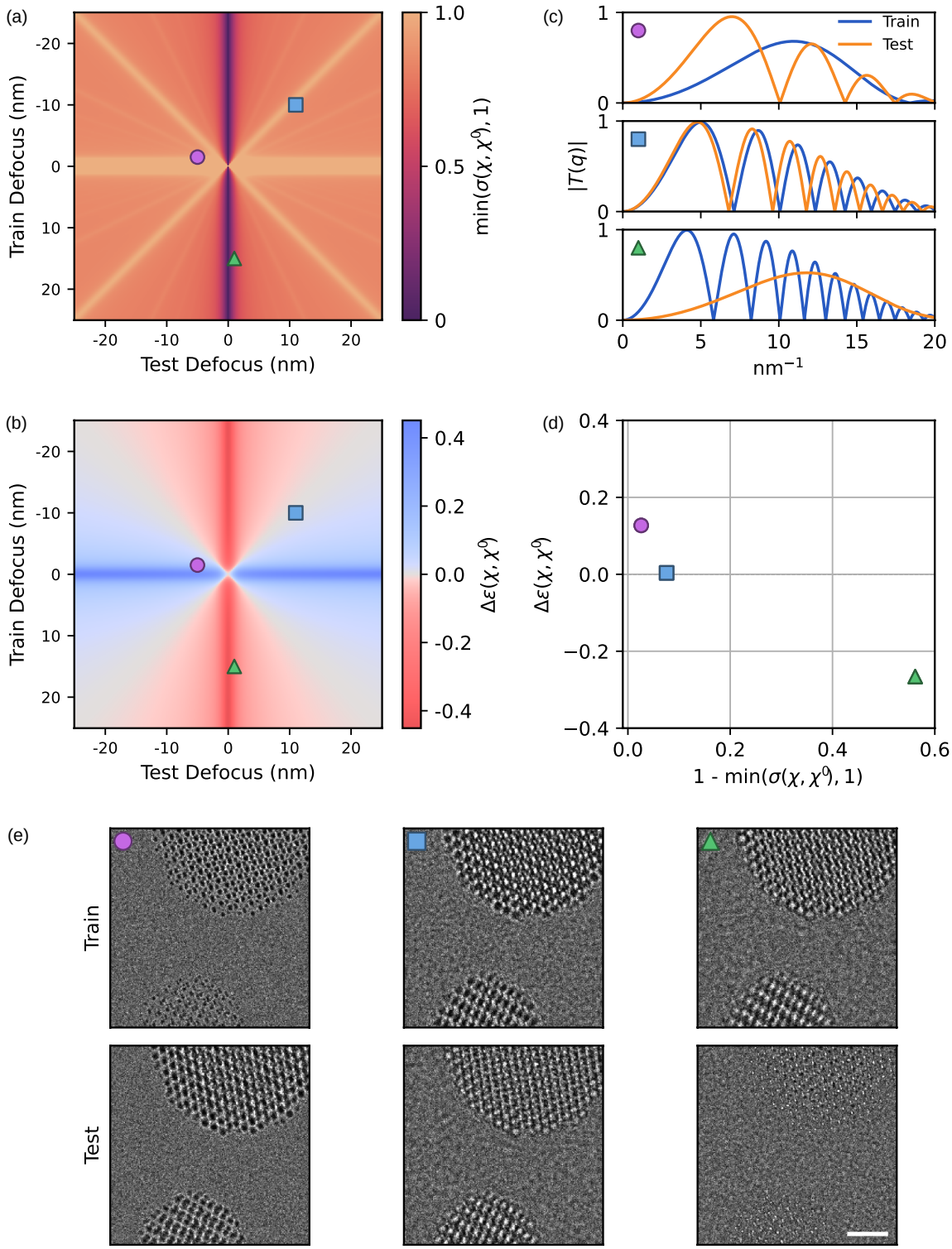


Fig. 3. Comparative visualization of (**a**) $\sigma(\chi, \chi')$ and (**b**) $\Delta\epsilon(\chi, \chi')$ for pairs of imaging conditions differing only in defocus. (**c**) Selected pairs of absolute contrast transfer functions (CTFs) corresponding to defocus conditions indicated by circle, square, and triangle markers. (**d**) Scatter plot demonstration of how these transfer functions relate to each other via the metrics $\sigma(\chi, \chi')$ and $\Delta\epsilon(\chi, \chi')$. (**e**) Example image pairs of CdSe nanoparticles on amorphous carbon matching selected CTF pairs. Imaging conditions correspond to a TEM operated at 300 kV with a focal spread of 10 Å. Scalebar is 1 nm.

be large in magnitude, indicating significant differences in the amount of total information available between two datasets.

For converged, numerically accurate calculation of the measures ϵ and σ, it is important to choose both appropriately large integration windows and fine sampling resolutions. In this study, for the aberration ranges chosen, a radial integration window between 0 and 30 inv. nm with a sampling resolution 0.05 nm were sufficiently accurate. Larger integration windows should be chosen for envelopes with slower decay, and finer resolutions should be chosen when comparing

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

CTFs with significant oscillatory behavior, e.g., when comparing aberrations with large defocus values. To improve numerical instability when comparing CTFs with very small aberration constants, we additionally add a small constant background (here, 1×10^{-6}) to the quantities $|T(\mathbf{q})|$.

Training Protocol and Hyperparameters

Our training and evaluation routines are written in PyTorch. All neural networks were optimized against a training dataset of 512 images before dihedral augmentation and a validation dataset of 128 images without augmentation. Our neural network architecture follows previous work: we utilize a U-Net with a ResNet backbone (Rangel DaCosta et al., 2024; Ronneberger et al., 2015), containing approximately 14 million trainable parameters. A visualization of the model architecture is included in the supplementary material (Supplementary Fig. 1). We use a cross-entropy loss function for training and evaluation, which, as implemented in PyTorch (Ansel et al., 2024), is

$$\mathcal{L}_{\mathrm{CE}}(\hat{y}, y) = -\sum_{c} \log\left[\frac{\exp(\hat{y}_c)}{\sum_i \exp(\hat{y}_i)}\right] y_c, \tag{9}$$

where the sum is taken over the segmentation classes—in this case, just two terms comprising the "nanoparticle" or "substrate" classes—and $\hat{y}_C$ and y_C are the predicted and true probabilities that a given pixel is the class C, respectively. Equation 9 has a range of $\left[-\log\left(\frac{e}{e+N-1}\right), -\log\left(\frac{1}{e+N-1}\right)\right]$ for N classes; we adjust the loss values in our analysis by the minimum attainable cross-entropy loss corresponding to a perfect two-class prediction ($\mathcal{L}_{\mathrm{CE}} \approx 0.3133$), so that all loss values fall between 0 and 1, i.e., $\tilde{\mathcal{L}}_{\mathrm{CE}}(\hat{y}, y) = \mathcal{L}_{\mathrm{CE}}(\hat{y}, y) + \log\frac{e}{e+1}$.

We trained two major sets of neural networks. Our first set of neural networks, comprising 3,520 total models, focuses on understanding OOD generalization dynamics across different hyperparameter conditions. These models were optimized for 128 epochs with stochastic gradient descent using four different learning rates of 5×10^{-4}, 2.3×10^{-3}, 1.07×10^{-2}, and 5×10^{-2}, four different batch sizes of 8, 16, 24, and 32 images, and a constant learning rate decay factor of 0.98. For each hyperparameter condition, we train 20 models. Throughout training, neural network performance was measured on a series of CdSe defocus datasets, described in the following section.

To further probe OOD behavior with respect to perturbations in imaging conditions, we train and evaluate a second series of models against training datasets with randomly sampled imaging conditions, this time with fixed hyperparameter conditions. The models were again trained with stochastic gradient descent with a learning rate of 1×10^{-2}, a batch size of 16, and a constant learning rate decay factor of 0.99. Models were optimized for up to a maximum of 32 epochs; we saved model weights at their best validation loss and after passing validation losses of 0.1 and 0.05, after which we conclude training. An additional 4,096 models were trained for these series; 3,745 models passed the 0.1 loss threshold and 3,089 models passed the 0.05 loss threshold. Only a small number of models surpassed the 0.025 loss threshold. After training, models were evaluated on their OOD performance against up to 32 randomly sampled evaluation datasets, which, after accounting for training and evaluation failures (primarily due to computation time outs), resulted in 80,896 and 72,864 total evaluations for the 0.1 loss 0.05 loss models, respectively.

Dataset Generation and Descriptions

Our segmentation dataset generation pipeline follows previous work (Rangel DaCosta et al., 2024). Throughout the structure and image generation process, we randomly sample structural configurations and imaging conditions. Unless specified otherwise, the sampling of parameters should be understood to be performed with uniform sampling throughout the range of a parameter. We partition training and validation dataset splits by structure, i.e., all models, per series, were trained on a single set of common structures, and validation performance (used to quantify both ID and OOD performance) was measured with a unique, disjoint set of structures; the training and validation structures are sampled via the same random structure generation protocol using equivalent parameters. We will begin by first discussing how the atomic structures themselves are sampled, then proceed with our multislice simulation protocol, and lastly, our approach for generating the final images and sampling imaging conditions via the sampling of aberration coefficients.

Nanoparticle Structure Generation and Sampling

We generate structures of nanoparticles on substrates using the Construction Zone software package and perform two simulations for each: one of the full structure, and one of just the nanoparticles in vacuum. Each simulation structure contains between one and five spherical nanoparticles with radii between 1 and 2.5 nm, separated by a distance of at least 1nm. We embed each particle atop an amorphous carbon substrate (Rangel DaCosta et al., 2024; Ricolleau et al., 2013) and at uniformly random orientations; the substrates have 12.8 nm square areas and thicknesses between 2 and 8 nm. The carbon substrate generation is further constrained to also satisfy periodic boundary conditions in the planar directions to prevent boundary artifacts in the TEM simulation. We add between 0 and 1 stacking faults to particles larger than 3 nm in diameter and between 0 and 2 stacking faults to particles larger than 4.4 nm. All results shown in the main manuscript utilize images of CdSe nanoparticles on carbon substrates. In Supplementary Section 2, we describe the generation of non-CdSe structures, which are used only for additional results shown in the Supplementary Material that are touched upon briefly in the Discussion.

Multislice Simulation

HRTEM simulations were performed with the multislice algorithm (Rangel DaCosta et al., 2021; Madsen & Susi, 2021), using a potential sampling of 0.1 Å, a slice thickness of 2 Å, eight frozen phonons, and infinitely projected potentials with the Kirkland potential parameterization (Kirkland, 2020). The final images are generated from the full structure simulations, while the segmentation labels are calculated from the electron wavefunction phase of the nanoparticles-in-vacuum simulation, applying a constant phase threshold of $\pi/16$ to identify the particle region alongside slight Gaussian blurring to smoothen the segmentation mask. We calibrate zero defocus to the minimum contrast condition of the nanoparticles in vacuum, which we heuristically determine via a two stage grid search to within 0.1 Å. After application of imaging conditions and noise effects, we randomly crop images into patches of 512 by 512 pixels and

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

Table 1. Maximum Magnitudes of Sampled Aberration Coefficients for Synthetically Generated HRTEM Images.

Aberration	Coefficient	Sampling Magnitude
Defocus	C_{20}	15 nm
Twofold astigmatism	C_{22}	3 nm
Axial coma	C_{31}	0.2 μm
Threefold astigmatism	C_{33}	0.2 μm
Spherical	C_{40}	0.1 mm

Table 2. Summary of Training Datasets, Corresponding Models, and in Which Figures Data are Utilized for Results and Analysis.

	Dataset A	Dataset B
Models trained	3,520	4,096
Aberrations	11 fixed defoci	Random, up to spherical
Hyperparameters	20	1
OOD evaluations	10	≤ 32
Relevant figures	4, 5, S4-6 S12, S13	6, S7-9, S11

then generate image augmentations using all eight possible dihedral augmentations arising from 90 degree rotations and flips along horizontal and vertical axes.

Image Augmentation and Aberration Sampling

We make two main sets of training datasets, one varying only in the image defocus (Dataset A), and the other including aberrations up to spherical aberration (Dataset B). For all images, we apply an electron dose and focal spread of 300e-/Å^2 and 10 Å, respectively. The defocus datasets consisted of imaging conditions sampled around focal points from −25 to +25 nm in steps of 5nm and with a standard deviation of ±1 nm. In Table 2, we summarize the sampling details used for each dataset, how models were trained and evaluated when using each dataset, and in which figures training results from each dataset are shown.

For our second series of datasets (Dataset B), we sample imaging conditions and generate image batches on-the-fly during model training and evaluation. For each individual image batch, we sample a target imaging condition by uniformly sampling aberration magnitudes up to a set of maximum values, given in Table 1, which are within correctable magnitudes for modern aberration-corrected instruments (Spence, 2017; Overwijk et al., 1997; Uhlemann & Haider, 1998; Lupini et al., 2010; Ophus et al., 2016). The magnitudes were chosen to depict an experimental setting in which the imaging conditions may vary but could be relatively controlled, and were tuned heuristically in order to improve the downstream dispersion of the sampled ($\Delta\epsilon$, σ) pairs. Within each image batch, different images will have slightly different imaging aberrations, too. We sample aberrations for each image from Gaussian distributions with means given by the previously sampled target imaging condition and the variances given as 1/8th of the maximum magnitudes. Aberration rotations were sampled with magnitudes up to 0.125 radians and variances of $\pi/16$ radians.

Within this second series, we additionally sample training dataset aberrations along a fixed combinations of defocus and spherical aberration to improve the dispersion of sampled aberration pairs with respect to ϵ and σ. For these aberrations, we fix defocus and spherical aberration via the following equation

$$C_{20} = -\sqrt{n\lambda C_{40}} \tag{10}$$

where λ is the electron wavelength. We define n as the passband order (differing in notation from the standard formulation used to analytically determine the Scherzer focus condition Spence, 2017). We sample 64 target aberrations along eight different passbands with $n = k/4$ for k from 1 through 6, 9, and 13. The 64 spherical aberration coefficients are linearly spaced between $0.1\,\mu\text{m} \leq C_{40} \leq C_{40,\max}$. We adjust the maximum magnitude of C_{40} so that the magnitudes of both defocus and spherical aberration remain smaller than the magnitudes indicated in Table 1, setting $C_{40,\max} = \min(0.1\,\text{mm}, (C_{20}^2/(\lambda n))$. As before, we sample the remaining aberrations uniformly at random but with reduced maximum magnitudes, scaled by a multiplicative factor of 0.2.

Results

Segmentation Network Training and Generalization Dynamics across Imaging Defoci

We begin by investigating the OOD generalization behavior solely with respect to image defocus in HRTEM, a relatively constrained setting which is nonetheless challenging to assess experimentally due to difficulties in precisely measuring and calibrating defocus. Moderate shifts in defocus can induce strong, nonlinear contrast effects which significantly alter image textures. At especially large defocus, image delocalization effects (Williams & Carter, 2008) and rapid oscillations of the contrast function make HRTEM micrographs especially hard to interpret and precisely measuring nanoparticle boundaries difficult. Altogether, these effects can impact the performance of neural networks used to segment atomic resolution HRTEM images. To help visually demonstrate how model performance degrades, and to provide some quantitative intuition for how different magnitudes of cross-entropy loss relate to final segmentation quality, we include a few example images of model segmentation in Figure 1. In section 1 of the Supplementary Material, we include more examples—including during earlier training stages—as well as distributional plots which demonstrate the relationship between the cross-entropy loss and the model Dice score, another common segmentation quality metric (Supplementary Figs. 2 and 3, respectively).

Utilizing a synthetic data-generation framework, we generate large simulated datasets of HRTEM images of CdSe nanoparticles on amorphous carbon substrates and train an ensemble of neural network models on datasets with defocus ranging from −25 to +25 nm in steps of 5nm. We calibrate defocus to the minimum contrast condition of the nanoparticles. In total, across the eleven defocus conditions, we train 3520 different models using 16 different hyperparameter settings, training 20 different randomly initialized models for each configuration to fully capture any distributional performance behavior, and at each epoch in training, we evaluate neural network performance on all defocus datasets. In Figure 4, we visualize ID (validation) and OOD training curves for models trained on −10, 0, and +10 nm defocus datasets across a selection of hyperparameter configurations and compare the trends and dynamics of neural network generalization. OOD

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

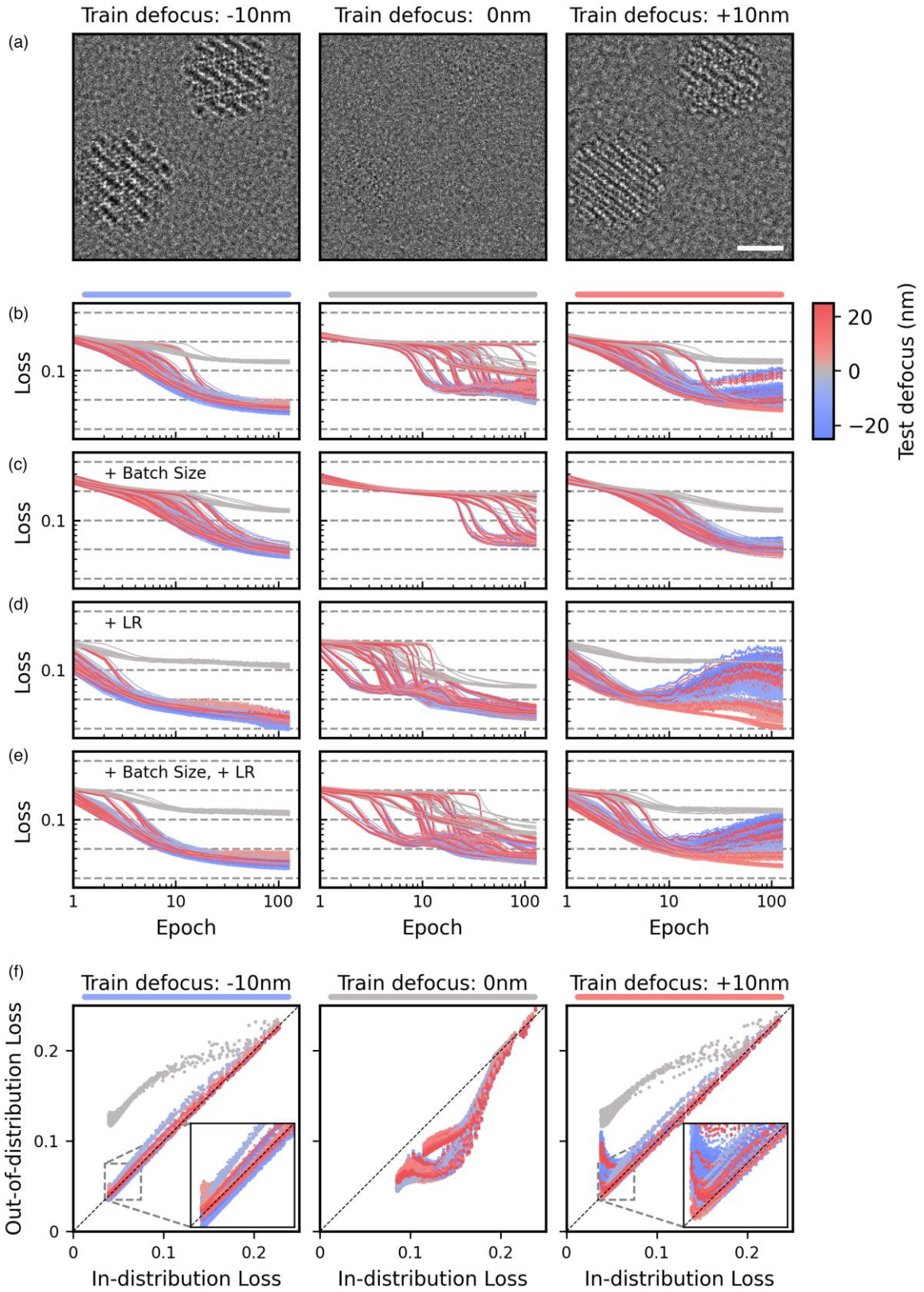


Fig. 4. Out-of-distribution generalization behavior throughout training of U-Net image segmentation models trained on −10 nm (left column), 0 nm (middle column) and +10 nm (right column) defocus images of CdSe nanoparticles. Defocus is measured from the minimum contrast condition of the CdSe nanoparticles. (**a**) Example regions from a simulated micrograph at each corresponding defocus for visual reference; scale bar is 1 nm. (**b**–**e**) Learning curves for series of models trained with learning rates of 2.3×10^{-3}, 2.3×10^{-3}, 1×10^{-2}, 1×10^{-2} and alternating batch sizes of 16, 32, 16, 32; line color corresponds to nominal defocus of validation dataset. (**f**) Correlation between in-distribution and OOD losses for models trained with learning rates of 1×10^{-2} and batch size of 16 (**d**), aggregated across all training times.

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

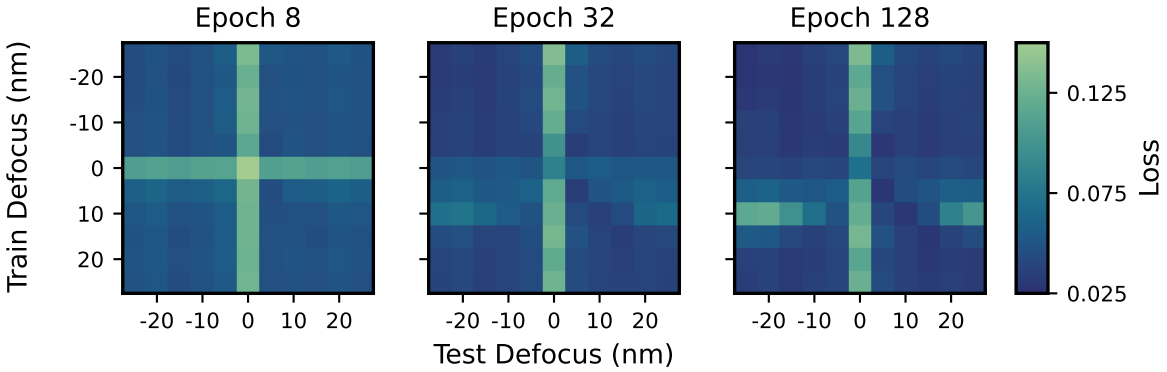


Fig. 5. Confusion matrix visualization of neural network loss as trained and evaluated on CdSe image datasets simulated with mean defoci between −25 and +25 nm at 300 kV and with a defocus spread of 10 Å. From left to right, we visualize model performance at epochs 8, 32, and 128. Visualized loss is taken as mean of twenty models trained with a learning rate of 1×10^{-2} and batch size of 24. Smaller loss is better.

performance tends to correlate with in-distribution performance (Fig. 4f), even though OOD performance is often slightly worse throughout training. The extent of performance degradation seems to depend slightly on the relative shift in defocus magnitude. In general, the 0 nm defocus dataset appears to be significantly more difficult for the neural networks to predict and learn from, as indicated by overall slow convergence rates of the 0 nm training curves and the high loss at the end of training. However, almost universally, the 0 nm-trained models are more successful OOD than they are in-distribution. For the +10 nm-trained models, there appears to be a significant onset of overfitting as indicated by an increase in loss and variance in performance at later stages of training (c.f. Figs. 4b, 4d, 4e, right).

The choice of training hyperparameters does not seem to significantly affect absolute performance or generalization behavior, but do seem to impact the training dynamics within a fixed training budget. Increasing the learning rate causes all models to reach their asymptotic optima earlier (c.f. Figs. 4b, 4d, 4e), but also leads to the models training on the +10 nm defocus dataset to enter deeper into an overfitting regime. Alternatively, increasing the batch size slightly slows down model convergence (c.f. Figs. 4b, 4c, 4e) but, for the nonzero defocus datasets, seems to reduce the variance in model performance and lessen the degree of overfitting. For the 0 nm-trained models, there is significantly higher variance on the dynamics of training, which may be an effect of decreased training stability due to the difficulty of the training dataset. Generalization dynamics trends are qualitatively similar across the remaining training defoci, for which we include results in Supplementary Figs. 4 and 5.

To further disentangle the relationship between training dataset defocus and neural network OOD performance across all defoci, we take a more detailed view of a single hyperparameter configuration of our models and explore performance across a series of training snapshots. We focus on three training stages—early in training, before models reach asymptotic convergence; near the optimum, as models start converging; and late in training, when models may begin overfitting. Using confusion matrices, in Figure 5, we visualize mean neural network OOD performance across all networks trained with a batch size of 24 images and a learning rate of 1×10^{-2} (corresponding to Fig. 4e). We highlight performance after the 8th, 32nd, and 128th (final) training epochs, corresponding to the three training stage described prior. Model performance during early phases of training seems relatively homogeneous across defocus for all models and performance OOD does not notably differ from in-distribution, except, again, for models trained with the "in-focus" 0 nm defocus dataset. After model performance begins to converge, OOD performance qualitatively settles, and may get slightly worse as training continues to progress, which can be seen when comparing performance of models trained on +5 nm and +10 nm defocus at epochs 32 and 128.

Model performance is best OOD when models evaluate test datasets containing images captured with similar defoci. Surprisingly, given the drastic image changes which can occur from contrast reversal effects, our models still perform about as well (but slightly worse, c.f. Supplementary Fig. 6) on defoci similar in magnitude but with opposite sign. Importantly, there is a significant asymmetry in the OOD performance trends between the training and test dataset pairs. When not overfitting, model performance on the 0 nm defocus dataset is usually the worst out of all ID and OOD datasets, across both different training defoci and training stages, improving only slightly during later stages in training. Yet, when training on the 0 nm defocus data, the models have the best OOD generalization—managing to perform better OOD than ID—while failing to fully converge during training. Models trained on slightly positive defocus had the worst generalization and most susceptibility to overfitting, which might be due to the intrinsic asymmetry of the atomic structures with respect to the electron beam—as simulated, the nanoparticles are always the first structure with which the electron beam interacts.

Use of Contrast Transfer Functions to Characterize OOD Generalization

The performance across defocus, and especially the performance on the 0 nm defocus datasets, can be partially understood by considering the relationship between contrast transfer functions (CTFs) of the two datasets. We introduce two auxiliary metrics, detailed in the Methods section, for comparing different imaging conditions by their CTFs: $\epsilon(\chi)$, a measure which could be interpreted as the amount of information transferred under a certain imaging condition, and $\sigma(\chi, \chi')$, which is an asymmetric measure of the relative information transfer overlap between two imaging conditions. The auxiliary metrics $\sigma(\chi, \chi')$ and $\epsilon(\chi)$ let us readily expand our analysis

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

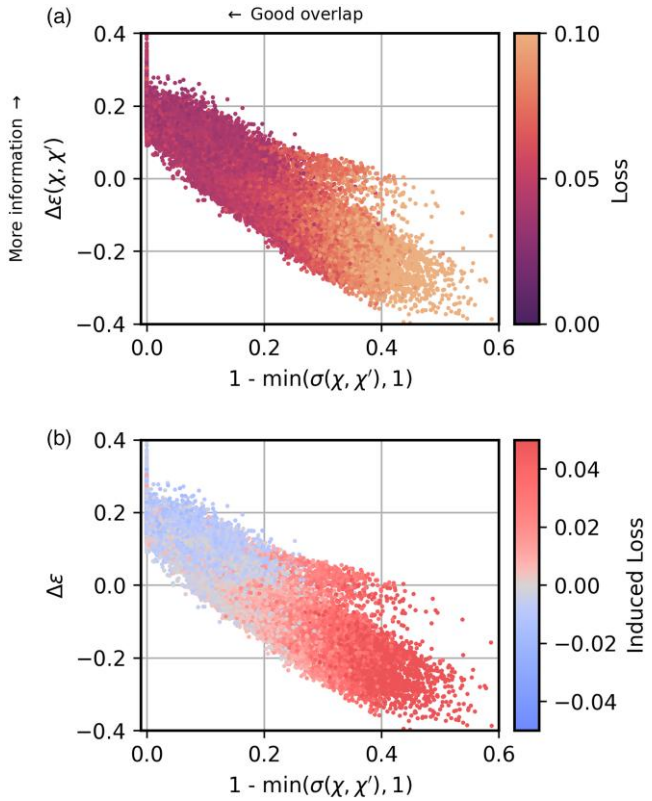


Fig. 6. Neural network loss (**a**) and induced (excess) loss over the training dataset (**b**) OOD, relative to $\sigma(\chi, \chi')$ and $\Delta\epsilon(\chi, \chi')$. Relative information content increases vertically, and information overlap decreases to the right. More negative induced loss (blue) indicates a model improving OOD, whereas a more positive induced loss (red) indicates the model worsening. OOD evaluations were performed using randomly generated model training datasets and evaluation datasets. Aberrations of training and evaluation datasets were sampled up to fifth order (defocus, axial coma, twofold and threefold astigmatism, and spherical), and models were trained until reaching a loss threshold of 0.05. Each marker corresponds to a unique training dataset–OOD dataset aberration pair, in total encompassing 3089 unique models and 72,864 unique dataset pairs.

beyond the simple case of varying just defocus to any arbitrary sets of imaging conditions.

Under this framework, we have performed another large ensemble neural network training series, training over 3,000 new models on randomly sampled imaging conditions with up to fifth-order aberrations. Each model is trained on unique, independently sampled imaging conditions. The models are trained for a fixed training period, and we save model weights at the optimal training epoch, as well as after models have passed loss thresholds of 0.1 and 0.05. After training, for each model and weight configuration, we sample up to 32 new imaging conditions and evaluate the model performance OOD, including over 72,000 unique model OOD evaluations from 3,089 total models. In Figure 6, we organize the OOD evaluations by their relative change in information quantity ($\Delta\epsilon(\chi, \chi')$) and overlap ($\sigma(\chi, \chi')$) between the training dataset and test datasets. Model loss OOD smoothly varies across differences in information content and overlap (Fig. 6a). When $\sigma(\chi, \chi')$ is close to 1—indicating strong overlap to the OOD imaging conditions—and $\Delta\epsilon(\chi, \chi')$ is small, neural network OOD performance does not differ strongly from ID performance. For large, positive $\Delta\epsilon(\chi, \chi')$, i.e., when the OOD dataset has more information transferred from the nanoparticles to the final image, the models can even perform better OOD than on their own training data (Fig. 6b). On the other hand, when $\Delta\epsilon(\chi, \chi')$ is large and negative, or when $\sigma(\chi, \chi')$ is small, OOD performance suffers, worsening as $\sigma(\chi, \chi')$ continues to decrease. Naturally, large negative values of $\Delta\epsilon$ can only occur when $\epsilon(\chi)$ itself is large, i.e., when the training dataset imaging conditions impart a significant amount of information transfer. Such datasets are in fact demonstrably easier to learn from: absolute model convergence and training convergence rates on in-distribution data both reliably improve with increasing $\epsilon(\chi)$ (c.f. Supplementary Fig. 7) indicating that—perhaps intuitively—neural network training dynamics strongly depend on the amount of information available in training datasets. Overall, there appears to be a zone of stability with large overlap and positive $\Delta\epsilon(\chi, \chi')$ where the neural network models perform comparatively well OOD, relative to their performance in-distribution. To help characterize the predictive power and relative uncertainty of these measurements, we include in the Supplementary Material visualizations of (induced) loss surfaces across different loss values and throughout a larger domain of $(\Delta\epsilon, \sigma)$, using Bayesian adaptive spline surfaces as a predictive model (c.f. Supplementary Fig. 8) (Friedman, 1991; Francom & Sansó, 2020), as well as quantitative estimates of the local variance in model OOD performance for the data shown in Figure 6 (c.f. Supplementary Fig. 9).

Discussion

Our main result presented in Figure 6 indicates that, for HRTEM image segmentation tasks, neural networks robustly handle small shifts in imaging conditions. Under larger perturbations to experimental imaging conditions, neural networks seem to gracefully and smoothly fail, and the extent of their failure can be directly understood by considering changes in the contrast transfer function of the microscope. Generally, this gives us confidence that neural network models for TEM analysis can be safely and meaningfully deployed in environments with (relatively controlled) shifting imaging conditions, and that we can rely on well-developed theories of imaging in electron microscopy to guide future development of machine learning approaches. Our results further indicate the possibility of designing training datasets which enable neural networks to safely generalize within a desired target domain. By improving the diversity of imaging conditions, the portion of the experimental domain to which the model cannot extrapolate shrinks: there are fewer possible examples which cannot be successfully analyzed. Thus, to control the amount of model degradation, we can try to guarantee that any new images encountered by the model have similar contrast transfer conditions to some portion of the training dataset—in practice, bounding the possible magnitudes of $\Delta\epsilon$ and σ for the data one reasonably expects to encounter by sampling sufficiently diverse imaging conditions during training.

For practical deployment of neural networks in electron microscopy, it is prudent to consider the limitations and possible extensions of our information transfer framework. We believe our approach can be readily adapted to other TEM imaging modalities, but it is not immediately clear that it would be appropriate across all types of machine learning models and tasks, e.g., regression with known differences in known OOD behavior (Ahuja et al., 2021). OOD generalization

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

behavior for supervised learning for classification might behave similarly given the similarities in construction between classification and segmentation tasks (Hooper et al., 2023). We expect our framework to be readily applicable to neural network models with different architectures or numbers of parameters, likely showcasing similar qualitative OOD behavior, given that the OOD shift is distinctly a data phenomenon. The degree of OOD degradation, however, may differ significantly: for example, for state-of-the-art architectures like vision transformers (Dosovitskiy et al., 2021; Paul & Chen, 2021), the measured loss surface may showcase larger regions of stable performance. Because our framework depends heavily on understanding precise relationships between training and test datasets, we it to be less useful for analyzing pretrained foundation models, such as the segment anything model (SAM) (Kirillov et al., 2023; Abebe et al., 2025), as information describing their training datasets can be quite scarce, especially for proprietary models.

When training neural network models for TEM analysis with simulated data, it is particularly important to understand any remaining gaps between simulated training data and experimental test data. For TEM applications, these emulation gaps largely involve describing and capturing representative imaging conditions and accurately accounting for the distribution of possible atomic structures. In this work, we only consider shifts in the imaging conditions due to shifts in the aberration conditions and high-frequency damping from chromatic aberration. We expect that detector effects and their corresponding modulation transfer functions, which are simple to model and incorporate into HRTEM simulation (Lee et al., 2014; Madsen et al., 2018) but more difficult to measure, could directly be incorporated into equation (3) as another multiplicative term. A fuller treatment of the envelope function (c.f. equation (4)) can be performed by incorporating the contributions from the partial coherence of the source, sample drift and vibration, and the detector (Spence, 2017; Williams & Carter, 2008). No further alterations to the metrics $\epsilon(\chi)$ and $\sigma(\chi, \chi')$ need to be made when experiments differ in their envelope functions. However, while differences in $\epsilon(\chi)$ remain bounded between -1 and 1 when multiple envelope functions are considered, the relative magnitudes of $\Delta\epsilon(\chi, \chi')$ and $\sigma(\chi, \chi')$ would shift, e.g., when the envelopes differ significantly in their high frequency cutoffs, which might make consistent interpretation of the two metrics across all imaging conditions more difficult. Lastly, while our framework is built upon the (weak) phase object model and utilizes no *a priori* information about the atomic structures, it seems to reasonably capture and describe OOD performance within a more experimentally realistic scenario (i.e., using multislice-simulated data). Understanding the limitations of the phase-object model, and how to extend or adapt a framework such as ours to more realistic theoretical settings, could be important for many experiments; a more rigorous analytic comparison of the information content in HRTEM data could utilize, for example, recent theoretical developments which describe the structural information captured in TEM data via Fisher information matrices (Dwyer, 2023; Dwyer & Paganin, 2024), which are readily adaptable to such settings and other experimental techniques (e.g., 4D-STEM).

Shifts in the distribution of the atomic structures are significantly more difficult to describe and anticipate in experiments —and, perhaps, more pressing to consider. Empirical evidence indicates that we can expect to maintain some degree of performance on OOD structures without immediately facing catastrophic failures (Sytwu et al., 2024; Rangel DaCosta et al., 2024). Within our imaging-centric framework, we could possibly incorporate structural information to predict OOD performance under shifts of atomic structure distributions by introducing weighted versions of the transfer function in equations (7) and (8) by substituting $T(\mathbf{q})$ with $\tilde{T}(\mathbf{q}) = \omega(\mathbf{q})T(\mathbf{q})$, where $\omega(\mathbf{q})$ is some spectral weighting function calculated from the atomic structures, e.g., arising from the atomic structure factors. We ran additional training experiments in which we trained segmentation networks on nanoparticle lattice systems varying in chemistry and symmetry (c.f. Supplementary Section 2), and show that OOD performance does not smoothly nor consistently vary across different regions of information overlap and information differences when using spectral weights determined derived from lattice structure factors (Supplementary Fig. 10), even when controlling for differences in chemistry. While this negative result is perhaps to be expected—it is optimistic to expect this simple framework to explain OOD generalization across a data domain which is both expansive and difficult to describe —we hope to highlight the need for more work in relating shifts in atomic structures to OOD model performance in TEM analysis tasks, and ideally, via physics-basic mechanistic explanations. More broadly, we still need to develop suitable techniques for describing shifts in atomic structure distributions; recent work on describing the relative information content of atomistic training datasets for machine learning of interatomic potentials (Schwalbe-Koda et al., 2025) could be pertinent, but did not yield meaningful results in our testing.

We strongly emphasize that our study relies on a fairly precise description of our data domain and of the OOD shifts we consider, using information theoretic approaches specific to electron microscopy. As we tackle more complex tasks on more challenging data with machine learning, producing such precise descriptions may become increasingly difficult. Domain agnostic methods for understanding OOD generalization behavior could be especially useful in circumventing challenges in describing differences in the relative information content between datasets for more complex tasks and datasets. Such methods exist, e.g., for detecting OOD training during model deployment (Madras et al., 2020; Sharma et al., 2021; Li et al., 2025; Lee et al., 2018; Haroush et al., 2022) or for conditioning models to respond more stably to OOD shifts from their training domains (Ramé et al., 2022; Rame et al., 2022; Mahajan et al., 2021; Arjovsky et al., 2020). However, few, if any, domain-agnostic approaches exist which can successfully calibrates or predict OOD generalization behavior; the authors are not aware of such an approach. We also explore the utility of second-order (Hessian) information of the model, known to strongly relate to the Fisher information (Böttcher & Wheeler, 2024; Rame et al., 2022), and empirical measures of the Kullback–Leiber (KL) divergence between imaging datasets as two potential complementary and domain-agnostic approaches for quantitatively understanding the OOD behavior of neural networks. For datasets for which neural networks perform well OOD, the OOD eigenspectrum distributions of neural network loss Hessians evolve similarly to those of the training dataset (Supplementary Fig. 12). Similarly, the dominant subspaces of the loss Hessians align over the course of model optimization (Supplementary Fig. 13). Together, these indicate that loss landscape curvature could be a good predictive measure

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

of OOD performance, and more fundamentally, that understanding the dependence of the loss landscape across different data distributions could provide deeper insight to OOD generalization behavior. From the data perspective, large KL divergences between datasets, which signify large differences in information content, moderately correlates to larger shifts in performance OOD, but it seems difficult to predict whether a model will degrade or improve (Supplementary Fig. 11). For more details on these experiments, we refer the reader to Supplementary Sections 3 and 4. Further investigation of accurate, inexpensive approaches for measuring domain shift and predicting OOD generalization can help bridge critical gaps in our understanding of generalizable machine learning approaches.

Conclusion

To summarize, our results indicate that neural networks for HRTEM analysis enjoy meaningful and stable OOD generalization behavior with respect to shifts in imaging conditions. Using a data-centric perspective and domain-specific information theoretic models, we can predict in which scenarios our models might begin to fail in experimental environments and can understand how to effectively design our training workflows to avoid such failures. Significant challenges remain, however, in our ability to fully describe our data at a distributional level, especially regarding atomic structures, which could remain a major barrier for understanding OOD generalization phenomena in materials characterization tasks. Still, we believe that maintaining a perspective focused on understanding how information arises in our experimental data and how machine learning models make use of such information will be crucial for refining our understanding and use of machine learning tools in applied scientific settings.

Data Availability

The data underlying this article will be shared on reasonable request to the corresponding author.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1093/mam/ozag067.

Acknowledgments

Work at the Molecular Foundry was supported by the Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy User Facility using NERSC award BES-ERCAP 32753. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0021110.

Author Contributions Statement

L.R.D. conducted the experiments and analyzed the results; L.R.D. and M.C.S. wrote and reviewed the manuscript; and M.C.S. supervised the work and acquired the funding.

Financial Support

The current study hasn't received any fund from any organizations or institutions.

Conflict of Interest

No competing interest is declared.

References

Abebe W, Strube J, Guo L, Tallent NR, Bel O, Spurgeon S, Doty C & Jannesari A (2025). SAM-I-Am: Semantic boosting for zero-shot atomic-scale electron micrograph segmentation. *Comput Mater Sci* **246**, 113400. https://doi.org/10.1016/j.commatsci.2024.113400

Ahuja K, Caballero E, Zhang D, Gagnon-Audet J-C, Bengio Y, Mitliagkas I & Rish I (2021). Invariance principle meets information bottleneck for out-of-distribution generalization. *Adv Neural Inf Process Syst* **34**, 3438–3450. https://dl.acm.org/doi/10.5555/3540261.3540524

Ansel J, Yang E, He H, Gimelshein N, Jain A, Voznesensky M, Bao B, Bell P, Berard D, Burovski E, Chauhan G, Chourdia A, Constable W, Desmaison A, DeVito Z, Ellison E, Feng W, Gong J, Gschwind M, Hirsh B, Huang S, Kalambarkar K, Kirsch L, Lazos M, Lezcano M, Liang Y, Liang J, Lu Y, Luk CK, Maher B, Pan Y, Puhrsch C, Reso M, Saroufim M, Siraichi MY, Suk H, Zhang S, Suo M, Tillet P, Zhao X, Wang E, Zhou K, Zou R, Wang X, Mathews A, Wen W, Chanan G, Wu P & Chintala S (2024). Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, pp. 929–947, New York, NY, USA: Association for Computing Machinery.

Arjovsky M, Bottou L, Gulrajani I & Lopez-Paz D (2020). 'Invariant risk minimization.' arXiv, arXiv:1907.02893, preprint: not peer reviewed.

Belardi CK, Lee C-H, Wang Y, Lovelace J, Weinberger KQ, Muller DA & Gomes CP (2025). Improving multislice electron ptychography with a generative prior. *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 3662–3672. https://doi.org/10.1109/ICCVW69036.2025.00382

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F & Vaughan JW (2010). A theory of learning from different domains. *Mach Learn* **79**(1), 151–175. https://doi.org/10.1007/s10994-009-5152-4

Borshon IZ, Jabbari V, Kingston TA, Shahbazian-Yassar R & Yurkiv V (2025). Deep learning analysis of solid-electrolyte interphase microstructures in lithium-ion batteries. *Adv Mater Interfaces* **12**(21), e00558. https://doi.org/10.1002/admi.202500558

Böttcher L & Wheeler G (2024). Visualizing high-dimensional loss landscapes with Hessian directions. *J Stat Mech* **2024**(2), 023401. https://doi.org/10.1088/1742-5468/ad13fc

Chen Z, Jiang Y, Shao Y-T, Holtz ME, Odstrčil M, Guizar-Sicairos M, Hanke I, Ganschow S, Schlom DG & Muller DA (2021). Electron ptychography achieves atomic-resolution limits set by lattice vibrations. *Science* **372**(6544), 826–831. https://doi.org/10.1126/science.abg2533

Cohen-Wang B, Vendrow J & Madry A (2025). Ask your distribution shift if pre-training is right for you. *Transactions on Machine Learning Research*.

de Jong AF & Van Dyck D (1993). Ultimate resolution and information in electron microscopy II. The information limit of transmission electron microscopes. *Ultramicroscopy* **49**(1), 66–80. https://doi.org/10.1016/0304-3991(93)90213-H

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J & Houlsby N (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

Dwyer C (2023). Quantum limits of transmission electron microscopy. *Phys Rev Lett* **130**(5), 056101. https://doi.org/10.1103/PhysRevLett.130.056101

Dwyer C & Paganin DM (2024). Quantum and classical Fisher information in four-dimensional scanning transmission electron microscopy. *Phys Rev B* **110**(2), 024110. https://doi.org/10.1103/PhysRevB.110.024110

Ede JM (2021). Deep learning in electron microscopy. *Mach Learn Sci Technol* **2**(1), 011004. https://doi.org/10.1088/2632-2153/abd614

Francom D & Sansó B (2020). BASS: An R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces. *J Stat Softw* **94**(8), 1–36. https://doi.org/10.18637/jss.v094.i08

Friedman JH (1991). Multivariate adaptive regression splines. *Ann Stat* **19**(1), 1–67. https://doi.org/10.1214/aos/1176347963

Groschner C, Choi C & Scott MC (2021). Machine Learning pipeline for segmentation and defect identification from high-resolution transmission electron microscopy data. *Microsc Microanal* **27**(3), 549–556. https://doi.org/10.1017/S1431927621000386

Haroush M, Frostig T, Heller R & Soudry D (2022). A statistical framework for efficient out of distribution detection in deep neural networks. *International Conference on Learning Representations*.

Hooper S, Chen MF, Kamal Saab K, Bhatia K, Langlotz C & Re C (2023). A case for reframing automated medical image classification as segmentation. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 55415–55441.

Kang K, Setlur A, Tomlin C & Levine S (2024). Deep neural networks tend To extrapolate predictably. *The Twelfth International Conference on Learning Representations*.

Khan A, Lee C-H, Huang PY & Clark BK (2023). Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *NPJ Comput Mater* **9**(1), 85. https://doi.org/10.1038/s41524-023-01042-3

Kim J, Rhee J, Kang S, Jung M, Kim J, Jeon M, Park J, Ham J, Kim BH, Lee WC, Roh S-H & Park J (2025). Self-supervised machine learning framework for high-throughput electron microscopy. *Sci Adv* **11**(14), eads5552. https://doi.org/10.1126/sciadv.ads5552

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y, Dollár P & Girshick R (2023). Segment anything. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

Kirkland EJ (2020). *Advanced Computing in Electron Microscopy*, 3rd ed. Cham, Switzerland: Springer Nature.

Küçükoğlu B, Mohammed I, Guerrero-Ferreira RC, Ribet SM, Varnavides G, Leidl ML, Lau K, Nazarov S, Myasnikov A, Kube M, Radecke J, Sachse C, Müller-Caspary K, Ophus C & Stahlberg H (2024). Low-dose cryo-electron ptychography of proteins at sub-nanometer resolution. *Nat Commun* **15**(1), 8062. https://doi.org/10.1038/s41467-024-52403-5

Larsen MHL, Lomholdt WB, Valencia CN, Hansen TW & Schiøtz J (2023). Quantifying noise limitations of neural network segmentations in high-resolution transmission electron microscopy. *Ultramicroscopy* **253**, 113803. https://doi.org/10.1016/j.ultramic.2023.113803

Lee C-H, Zeltmann SE, Yoon D, Ma D & Muller DA (2025). PtyRAD: A high-performance and flexible ptychographic reconstruction framework with automatic differentiation. *Microsc Microanal* **31**(4), ozaf070. https://doi.org/10.1093/mam/ozaf070

Lee K, Lee K, Lee H & Shin J (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 7167–7177. https://dl.acm.org/doi/10.5555/3327757.3327819

Lee Z, Rose H, Lehtinen O, Biskupek J & Kaiser U (2014). Electron dose dependence of signal-to-noise ratio, atom contrast and resolution in transmission electron microscope images. *Ultramicroscopy* **145**, 3–12. https://doi.org/10.1016/j.ultramic.2014.01.010

Li P, Cao B, Zhang C & Hu Q (2025). 'Generalized few-shot out-of-distribution detection.' arXiv, arXiv:2508.05732, preprint: not peer reviewed.

Lin R, Zhang R, Wang C, Yang X-Q & Xin HL (2021). TEMImageNet training library and AtomSegNet deep-learning models for high-precision atom segmentation, localization, denoising, and deblurring of atomic-resolution images. *Sci Rep* **11**(1), 5386. https://doi.org/10.1038/s41598-021-84499-w

Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H & Cui P (2023). 'Towards out-of-distribution generalization: A survey.' arXiv, arXiv:2108.13624, preprint: not peer reviewed.

Liu Z, Xu Y, Xu Y, Qian Q, Li H, Jin R, Ji X & Chan AB (2022). An empirical study on distribution shift robustness from the perspective of pre-training and data augmentation. *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*.

Lobato I, Friedrich T & Van Aert S (2024). Deep convolutional neural networks to restore single-shot electron microscopy images. *NPJ Comput Mater* **10**(1), 10. https://doi.org/10.1038/s41524-023-01188-0

Lupini AR, Wang P, Nellist PD, Kirkland AI & Pennycook SJ (2010). Aberration measurement using the Ronchigram contrast transfer function. *Ultramicroscopy* **110**(7), 891–898. https://doi.org/10.1016/j.ultramic.2010.04.006

Ma D, Li G, Muller DA & Zeltmann SE (2026). Information in 4D-STEM: Where it is, and how to use it. *Ultramicroscopy* **283**. https://doi.org/10.1016/j.ultramic.2026.114351

Madras D, Atwood J & D'Amour A (2020). Detecting extrapolation with local ensembles. *International Conference on Learning Representations*.

Madsen J, Liu P, Kling J, Wagner JB, Hansen TW, Winther O & Schiøtz J (2018). A deep learning approach to identify local structures in atomic-resolution transmission electron microscopy images. *Adv Theory Simul* **1**(8), 1800037. https://doi.org/10.1002/adts.201800037

Madsen J & Susi T (2021). The abTEM code: Transmission electron microscopy from first principles. *Open Res Eur* **1**, 24. https://doi.org/10.12688/openreseurope.13015.1

Mahajan D, Tople S & Sharma A (2021). Domain generalization using causal matching. *Proceedings of the 38th International Conference on Machine Learning* **139**, 7313–7324.

McCray ARC, Ribet SM, Varnavides G & Ophus C (2025). Accelerating iterative ptychography with an integrated neural network. *J Microsc* **300**(2), 180–190. https://doi.org/10.1111/jmi.13407

Miller J, Taori R, Raghunathan A, Sagawa S, Koh PW, Shankar V, Liang P, Carmon Y & Schmidt L (2021). Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution deneralization. *Proceedings of the 38th International Conference on Machine Learning* **139**, 7721–7735.

Munshi J, Rakowski A, Savitzky BH, Zeltmann SE, Ciston J, Henderson M, Cholia S, Minor AM, Chan MKY & Ophus C (2022). Disentangling multiple scattering with deep learning: Application to strain mapping from electron diffraction patterns. *NPJ Comput Mater* **8**(1), 1–15. https://doi.org/10.1038/s41524-022-00939-9

Nellist PD, McCallum BC & Rodenburg JM (1995). Resolution beyond the "information limit" in transmission electron microscopy. *Nature* **374**(6523), 630–632. https://doi.org/10.1038/374630a0

Ophus C, Rasool HI, Linck M, Zettl A & Ciston J (2016). Automatic software correction of residual aberrations in reconstructed HRTEM exit waves of crystalline samples. *Adv Struct Chem Imaging* **2**(1), 15. https://doi.org/10.1186/s40679-016-0030-1

Overwijk MHF, Bleeker AJ & Thust A (1997). Correction of three-fold astigmatism for ultra-high-resolution TEM. *Ultramicroscopy* **67**(1), 163–170. https://doi.org/10.1016/S0304-3991(96)00096-4

Paul S & Chen P-Y (2021). Vision transformers are robust learners. *Proceedings of the AAAI Conference on Artificial Intelligence* **36**(2), 2071–2081. https://doi.org/10.1609/aaai.v36i2.20103

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026

Rame A, Dancette C & Cord M (2022). Fishr: Invariant gradient variances for out-of-distribution generalization. *Proceedings of the 39th International Conference on Machine Learning* **162**, 18347–18377.

Ramé A, Kirchmeyer M, Rahier T, Rakotomamonjy A, Gallinari P & Cord M (2022). Diverse weight averaging for out-of-distribution generalization. *Adv Neural Inf Process Syst* **35**, 10821–10836. https://dl.acm.org/doi/10.5555/3600270.3601056

Rangel DaCosta L, Brown HG, Pelz PM, Rakowski A, Barber N, O'Donovan P, McBean P, Jones L, Ciston J, Scott MC & Ophus C (2021). Prismatic 2.0—simulation software for scanning and high resolution transmission electron microscopy (STEM and HRTEM). *Micron* **151**, 103141. https://doi.org/10.1016/j.micron.2021.103141

Rangel DaCosta L, Sytwu K, Groschner CK & Scott MC (2024). A robust synthetic data generation framework for machine learning in high-resolution transmission electron microscopy (HRTEM). *NPJ Comput Mater* **10**(1), 1–11. https://doi.org/10.1038/s41524-024-01336-0

Ricolleau C, Le Bouar Y, Amara H, Landon-Cardinal O & Alloyeau D (2013). Random vs realistic amorphous carbon models for high resolution microscopy and electron diffraction. *J Appl Phys* **114**(21), 213504. https://doi.org/10.1063/1.4831669

Ronneberger O, Fischer P & Brox T (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* **9351**. https://doi.org/10.1007/978-3-319-24574-4_28

Rose H (1984). Information transfer in transmission electron microscopy. *Ultramicroscopy* **15**(3), 173–191. https://doi.org/10.1016/0304-3991(84)90038-X

Sadre R, Ophus C, Butko A & Weber GH (2021). Deep learning segmentation of complex features in atomic-resolution phase-contrast transmission electron microscopy images. *Microsc Microanal* **27**(4), 804–814. https://doi.org/10.1017/S1431927621000167

Schwalbe-Koda D, Hamel S, Sadigh B, Zhou F & Lordi V (2025). Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *Nat Commun* **16**(1), 4014. https://doi.org/10.1038/s41467-025-59232-0

Sharma A, Azizan N & Pavone M (2021). Sketching curvature for efficient out-of-distribution detection for deep neural networks. *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* **161**, 1958–1967.

Shi J, Gare G, Tian J, Chai S, Lin Z, Vasudevan A, Feng D, Ferroni F & Kong S (2024). LCA-on-the-line: Benchmarking out of distribution generalization with class taxonomies. *Proceedings of the 41st International Conference on Machine Learning* **235**, 44887–44908.

Spence JCH (2017). *High-Resolution Electron Microscopy*, 4th ed. Oxford, New York: Oxford University Press.

Sytwu K, Groschner C & Scott MC (2022). Understanding the influence of receptive field and network complexity in neural network-guided TEM image analysis. *Microsc Microanal* **28**(6), 1896–1904. https://doi.org/10.1017/s1431927622012466

Sytwu K, Rangel DaCosta L & Scott MC (2024). Generalization across experimental parameters in neural network analysis of high-resolution transmission electron microscopy datasets. *Microsc Microanal* **30**(1), 85–95. https://doi.org/10.1093/micmic/ozae001

Uhlemann S & Haider M (1998). Residual wave aberrations in the first spherical aberration corrected transmission electron microscope. *Ultramicroscopy* **72**(3), 109–119. https://doi.org/10.1016/S0304-3991(97)00102-2

Van Dyck D & de Jong AF (1992). Ultimate resolution and information in electron microscopy: General principles. *Ultramicroscopy* **47**(1), 266–281. https://doi.org/10.1016/0304-3991(92)90202-U

Varnavides G, Bekkevold JM, Ribet SM, McCray A, Scott MC, Jones L & Ophus C (2026). Beyond contrast transfer: Spectral SNR as a finite-dose metric for STEM phase retrieval. *Microsc Microanal* **32**(2). https://doi.org/10.1093/mam/ozag005

Varnavides G, Ribet SM, Zeltmann SE, Yu Y, Savitzky BH, Byrne DO, Allen FI, Dravid VP, Scott MC & Ophus C (2024). 'Iterative phase retrieval algorithms for scanning transmission electron microscopy.' arXiv, arXiv:2309.05250, preprint: not peer reviewed.

Vincent JL, Manzorro R, Mohan S, Tang B, Sheth DY, Simoncelli EP, Matteson DS, Fernandez-Granda C & Crozier PA (2021). Developing and evaluating deep neural network-based denoising for nanoparticle TEM images with ultra-low signal-to-noise. *Microsc Microanal* **27**(6), 1431–1447. https://doi.org/10.1017/S1431927621012678

Wang X, Li J, Ha HD, Dahl JC, Ondry JC, Moreno-Hernandez I, Head-Gordon T & Paul Alivisatos A (2021). AutoDetect-mNP: An unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. *JACS Au* **1**(3), 316–327. https://doi.org/10.1021/jacsau.0c00030

Williams DB & Carter CBarry (2008). *Transmission Electron Microscopy: A Textbook for Materials Science*, 2nd ed. New York: Springer.

Ye H, Xie C, Cai T, Li R, Li Z & Wang L (2021). Towards a theoretical framework of out-of-distribution generalization. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pp. 23519–23531, Red Hook, NY, USA: Curran Associates Inc.

Ye N, Li K, Bai H, Yu R, Hong L, Zhou F, Li Z & Zhu J (2022). OoD-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7937–7948. New Orleans, LA, USA: IEEE.

Yu H, Liu J, Zhang X, Wu J & Cui P (2024). 'A survey on evaluation of out-of-distribution generalization.' arXiv, arXiv:2403.01874, preprint: not peer reviewed.

Downloaded from academic.oup.com/mam/article/32/4/ozag067/8734916 by guest on 10 November 2026