IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 045003 https://doi.org/10.1088/2632-2153/abf9b1

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
23 December 2020

REVISED
16 March 2021

ACCEPTED FOR PUBLICATION
20 April 2021

PUBLISHED
15 July 2021

PAPER

Neural network analysis of neutron and x-ray reflectivity data: pathological cases, performance and perspectives

Alessandro Greco[1], Vladimir Starostin[1], Alexander Hinderhofer[1,*], Alexander Gerlach[1], Maximilian W A Skoda[2], Stefan Kowarik[3] and Frank Schreiber[1,*]

[1] Institute of Applied Physics, University of Tübingen, 72076 Tübingen, Germany
[2] ISIS Pulsed Neutron and Muon Source, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Science and Innovation Campus, Didcot OX11 0QX, United Kingdom
[3] Department of Physical Chemistry, University of Graz, 8010 Graz, Austria
[*] Authors to whom any correspondence should be addressed.

E-mail: alexander.hinderhofer@uni-tuebingen.de **and** frank.schreiber@uni-tuebingen.de

Keywords: thin films, reflectivity, data analysis, neural networks

Supplementary material for this article is available online

Abstract

Neutron and x-ray reflectometry (NR and XRR) are powerful techniques to investigate the structural, morphological and even magnetic properties of solid and liquid thin films. While neutrons and x-rays behave similarly in many ways and can be described by the same general theory, they fundamentally differ in certain specific aspects. These aspects can be exploited to investigate different properties of a system, depending on which particular questions need to be answered. Having demonstrated the general applicability of neural networks to analyze XRR and NR data before (Greco *et al* 2019 *J. Appl. Cryst.* **52** 1342), this study discusses challenges arising from certain pathological cases as well as performance issues and perspectives. These cases include a low signal-to-noise ratio, a high background signal (e.g. from incoherent scattering), as well as a potential lack of a total reflection edge (TRE). By dynamically modifying the training data after every mini batch, a fully-connected neural network was trained to determine thin film parameters from reflectivity curves. We show that noise and background intensity pose no significant problem as long as they do not affect the TRE. However, for curves without strong features the prediction accuracy is diminished. Furthermore, we compare the prediction accuracy for different scattering length density combinations. The results are demonstrated using simulated data of a single-layer system while also discussing challenges for multi-component systems.

1. Introduction

To investigate the structural, morphological or magnetic properties of surfaces and layered structures, such as solid and liquid thin films [1–8], x-ray and neutron reflectometry (XRR and NR) are often employed due to an array of benefits. Reflectometry measurements are an excellent non-invasive method for gaining access to the layer thickness, interface roughness and scattering length density (SLD) of a large variety of thin films [9]. Although neutrons and x-rays behave similarly in many ways, they also show some key differences regarding their elementary scattering process [10]. While x-rays interact with electrons, neutrons mainly interact with the nuclei (except for magnetic effects, which we neglect here), which allows them to be employed as probes for different types of samples, and thus answer different questions in a complementary manner [11]. Importantly, these differences are also reflected in the data the two methods produce and they must be taken into account during data analysis in order to extract the correct information from a given measurement [12].

The measured scattering signal is generally based on a Fourier transform of the probed structure in combination with Fresnel reflection coefficients; however, due to multiple scattering and the loss of the scattering phase in the detection process, it is not trivial to reconstruct the original real space structure, since a direct inverse transformation of the data is not possible. For reflectometry, a common way to extract

Table 1. Parameter ranges for the simulated training, validation and test data.

	Thickness (Å)	Roughness (Å)	SLD (10^{-6}Å^{-2})
Ambient	—	—	0
Layer	20–300	0–60	−8–16
Substrate	—	0–10	−8–16

information from the measured data is to use a recursive mathematical model [13–16] to simulate a reflected intensity profile $R(q)$ for different scattering vectors q which is in agreement with the measurement. This is usually done via iterative least mean squares fitting algorithms which have been implemented in various free software packages [17–21]. Increasingly sophisticated ways of optimizing the search for a local minimum are continuously developed to make the fitting process as fast and reliable as possible. However, depending on the quality of the data and the complexity of the studied system, finding a suitable model still often requires prior knowledge, a considerable amount of expertise, and is generally time-consuming.

A promising alternative could be machine learning (ML) techniques, which recently have been demonstrated for various scientific questions on related scattering techniques, such as small-angle scattering [22–24] and crystal structure or symmetry determination [25–27]. As shown in a recent study, mostly focused on real-time XRR [28], fully-connected neural networks can be trained to determine thickness, roughness and SLD parameters directly from the measured reflectivity data with very high speed and good accuracy within a comparatively large parameter range, thereby reducing the need for user input.

In this paper, we discuss the analysis of reflectivity data using neural networks in the light of three types of challenges: (1) Reflectivity curves without strong features that have a low information content, (2) curves without a total reflection edge (TRE), and (3) data with significant noise or background. We demonstrate that by applying different types of random noise and background intensity to the training data during the training process, the resulting neural network model is robust toward most types of perturbations when determining thin film properties, but struggles with certain particularly difficult edge cases. We test this approach on simulated reflectivity data of a single layer plus substrate within the same parameter ranges.

2. Methods

2.1. Reflectivity data simulation

The training and validation data were simulated using a model of a single layer on a substrate in air as an ambient medium. The model had five open parameters: substrate roughness, substrate SLD, layer thickness, layer roughness and layer SLD. We generated 3×10^6 parameter sets for training and additionally 2×10^4 sets for validation. The values of each set were generated within the ranges given in table 1 with a higher sampling density toward the limits of each range. This was done to make the local density of sampled values near the limits more similar to that toward the center of the distribution. The number of generated parameter sets was chosen as a compromise to cover as much of the large parameter space as possible while still maintaining technical feasibility in terms of training time and occupied memory. The range of possible SLD values for the substrate and layer was specifically designed to encompass a large spectrum of negative and positive SLDs of the most common elements. This allowed us to investigate the effects of different combinations of negative and positive SLDs on the prediction performance of the neural network. Furthermore, SLD combinations with contrasts between the layer and the substrate and the layer and the ambient SLD of less than 1×10^{-6} Å^{-2} were excluded from the training and testing data. These are known to produce curves without strong features and excluding them boosted the performance even on more feature-rich data. Also, for reasons of practicability due to the chosen q range, we focus on film thicknesses larger than 20 Å. A brief performance comparison between models trained with and without those exclusions is shown in figure S1 of the supporting information (available online at stacks.iop.org/MLST/2/045003/mmedia).

From the generated parameter sets, reflectivity curves were simulated using the implementation of the Matrix method [15] in the *refl1D* package [17]. The simulated q range was restricted to a range of 0.01–0.3 Å^{-1} in order to avoid q ranges where Bragg reflections and Laue oscillations might appear in real measurements since they are not described by our slab model. Thus, we can be sure that the neural network predictions are only based on Kiessig fringes and other features related to the layer structure which would be present in an experimentally measured curve. Within this q range, the reflected intensity values R_i were simulated at 100 equally-spaced discrete points q_i. This number was chosen to be comparable with common point densities of experiments. Of course experimental data would also be subject to a finite q resolution, however, to limit the number of noise sources under study, we have chosen to approximate this with uniform noise as described in section 2.1.2.

During training, different types of noise and background intensity were added to each curve every time a mini-batch was drawn from the training set. This means that every time the neural network encounters one of the training curves, the curve is modified with different noise and background. This step is crucial to avoid overfitting and to prevent the network from learning heuristics that do no work on imperfect data by forcing it to learn how to denoise the input data before interpreting it. The perturbations added to the data were Poisson noise, uniform noise, curve scaling and a constant background. Each type of curve modification is described in detail in the following.

2.1.1. Poisson noise due to counting statistics

Statistical noise in scattering data results from the counting statistics of scattered particles arriving at the detector and is dependent on the expected counting rate N, i.e. the recorded intensity. Since this noise generally follows a Poisson distribution, the noise of a simulated reflectivity curve can be calculated by replacing each intensity value R_i with a random value picked from the distribution

$$f_s(x) = \frac{f(x; sR_i)}{s} \tag{1}$$

where s is the theoretical maximum number of counts at total reflection (for a monochromatic experiment) and f is the Poisson distribution

$$f(x; N) = \frac{N^x e^N}{x!}. \tag{2}$$

Since the simulated intensities R only range from 0 to 1, they must be scaled to values which could occur in an experiment $N_i = sR_i$ before calculating the noise. In this study, for every curve a scaling factor (corresponding to the flux) was randomly chosen on a logarithmic scale within $s = [10^6, 10^8]$ to represent different experimental conditions. For the upper limit of $s = 10^8$, there is no noticeable noise in the chosen q range anymore. An example of a curve with a noise level of $s = 10^6$ is shown in figure 1(a).

2.1.2. Uniform noise

Since the statistical noise mainly affects low-intensity regions in a reflectivity curve, the first half of the curve, i.e. low-q features and in particular the TRE, remain mostly unaffected by it. Despite that, experimental data may contain noise or other small deviations in this region. For example, in time-of-flight (TOF) neutron scattering experiments, the intensity given by the normalized reflectivity curve is not necessarily proportional to the counts on the detector, since the incoming beam generally has an energy spectrum with a non-uniform distribution (e.g. Maxwell–Boltzmann). The final intensity of the curve is then obtained by normalizing the number of counts in each channel (i.e. q value) with the corresponding incoming intensity of that energy. This can effectively lead to worse counting statistics in regions with seemingly higher intensity, such as near the TRE, compared to lower intensity regions. Furthermore, the beam shape and random errors on the measurement angle typically lead to non-negligible deviations of the intensity. This is particularly pronounced near the TRE where slight errors on the angle might translate to large errors in intensity.

To make the neural network robust against these types of errors, a random, uniformly distributed scaling factor α_i is multiplied to each intensity value R_i of each input curve R, so that the new intensity is given by

$$R_i^* = \alpha_i R_i \tag{3}$$

where $\alpha = [0.7, 1.3]$. An example of a curve with uniform noise applied to it is shown in figure 1(b).

2.1.3. Curve scaling

In order to analyze reflectivity data, the measured intensity is typically normalized to the intensity at total reflection. For monochromatic experiments, this step is preceded by an angular dependent footprint correction. For polychromatic experiments (e.g. TOF), the normalization must in addition take into account the above mentioned energy spectrum, usually obtained via measuring the direct beam. Both of these corrections produce an error on the normalization procedure (which itself has a finite accuracy) and may result in distortions of the data. This effect is further exacerbated if there is no TRE, since the intensity at total reflection is not directly available, i.e. the naturally given absolute scale of the TRE is missing.

To make the neural network robust against these slight distortions, during training, every input curve R is multiplied by a random, uniformly distributed scaling factor β, so that the new curve R^* is given by

$$R^* = \beta R \tag{4}$$

where $\beta = [0.9, 1.1]$. An example of a curve with random scaling applied to it is shown in figure 1(b).

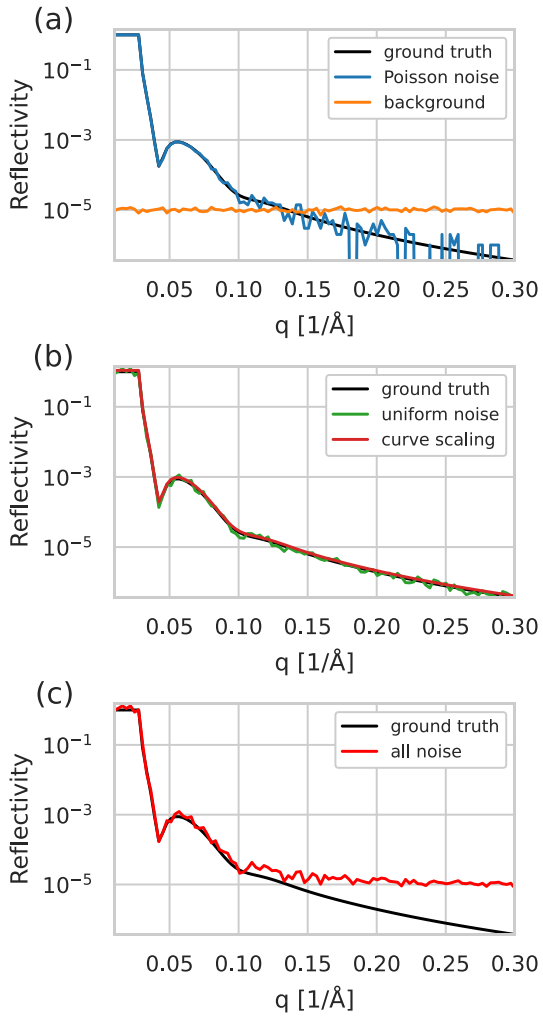

Figure 1. Example of a simulated reflectivity curve with different noise and background applied to it. (a) The ground truth curve and the same curves with Poisson noise with a level $s = 10^6$ as well as a background of $b = 10^{-5}$. (b) The same curves with curve scaling and uniform noise applied to it, respectively. (c) Comparison of the ground truth with a curve with all four modifications applied to it.

2.1.4. Residual background

Both x-ray and neutron scattering experiments contain background intensity stemming from various sources, such as background radiation or detector noise. Within the q range discussed in this study (max. 0.3 $Å^{-1}$), for XRR these effects are usually negligible compared to the measured intensity of the reflected beam.

In addition, neutron reflectivity data usually contains background resulting from incoherent scattering [10]. In practice, most of this background is already removed during the data reduction step, e.g. via calibration with a pure transmission measurement. During data analysis, the residual background is then routinely approximated by a constant value, although more complex models exist [12].

To account for this, the residual background in the data was approximated by a q-independent constant b with normally distributed fluctuations with a standard deviation of $\sigma_b = 0.1b$. The fluctuations were added so that the original curve cannot be fully reconstructed by just subtracting a constant value. Thus, for a given curve, the background added to each intensity value was randomly picked from the normal distribution

$$p_b(x; b, \sigma_b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{(x-b)^2}{2\sigma_b^2}\right). \tag{5}$$

In this work, the background level of each curve was randomly chosen on a logarithmic scale within $b = [10^{-7}, 10^{-4}]$. An example of an added background with $b = 10^{-5}$ shown in figure 1(a).

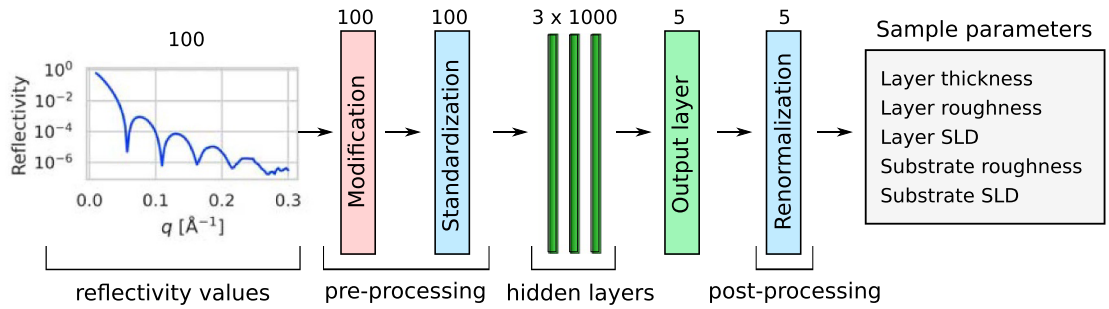

Figure 2. Architecture of the fully-connected neural network used in this study. The addition of noise and background to the input data during the modification step was only used during training.

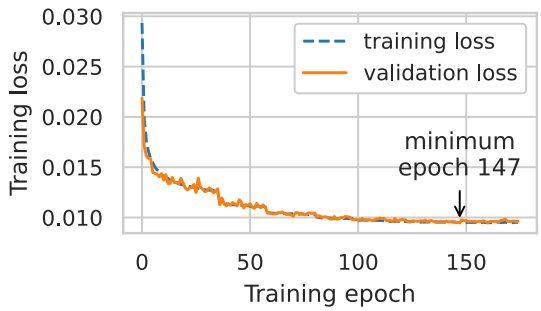

Figure 3. Training and validation loss during training. The training was stopped after 175 epochs because the validation error did not decrease further. The lowest validation error was observed at epoch 147.

2.2. Neural network design and training

The neural network used in this study was a fully-connected model with 100 input neurons, three hidden layers with 1000 neurons each and five output neurons as shown in figure 2. It was written in Python 3.7 with the help of the TensorFlow (2.1) framework [29]. The input corresponds to the reflected intensity values $R \in \mathbb{R}^{100}$ at 100 discrete points in q space as described in the previous section. The output corresponds to the five open thin film parameters $y \in \mathbb{R}^5$ as shown in table 1. As an activation function, a simple ReLU (rectified linear) unit was chosen for all layers. During training, whenever a mini batch of 512 curves is drawn from the training set, curve modifications are applied as described in section 2.1. Then, each input R_i is independently standardized by subtracting the mean $\bar{R}_i$ and dividing by the standard deviation $\tilde{R}_i$ of all values of that input in the entire, randomly modified training set. The standardized input is thus given by

$$\hat{R}_i = \frac{R_i - \bar{R}_i}{\tilde{R}_i} \tag{6}$$

with

$$\bar{R}_i = \frac{1}{N}\sum_{n=1}^{N} R_{n,i} \quad \text{and} \quad \tilde{R}_i = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(R_{n,i} - \bar{R}_i)^2} \tag{7}$$

where R_n is a curve from the training set of size $N = 3 \times 10^6$.

The output values y_j were normalized by the greater absolute value of either the minimum or maximum of their respective ranges given in table 1. This effectively confined all output values to a range from -1 to 1.

The ADAM algorithm [30] was used as an optimizer with the recommended default parameters and a starting learning rate of 10^{-3}. Furthermore, the learning rate was reduced by half each time the validation loss stagnated for ten epochs in a row in order to avoid skipping over narrow minima in the loss function space. The mean squared error (MSE) of the normalized outputs was used as the loss function. The neural network was trained on a GeForce RTX 2080 Ti GPU and an Intel® Core™ i5-9600K CPU for 175 epochs with a training time of about 6.5 min per epoch, amounting to a total training time of about 19 h. During training the training and losses were monitored as shown in figure 3. Overall, the training and validation loss were almost identical, with the validation loss only being slightly higher after 70 epochs. The reason why this difference is so small is that the training data is randomly modified with noise and background during every

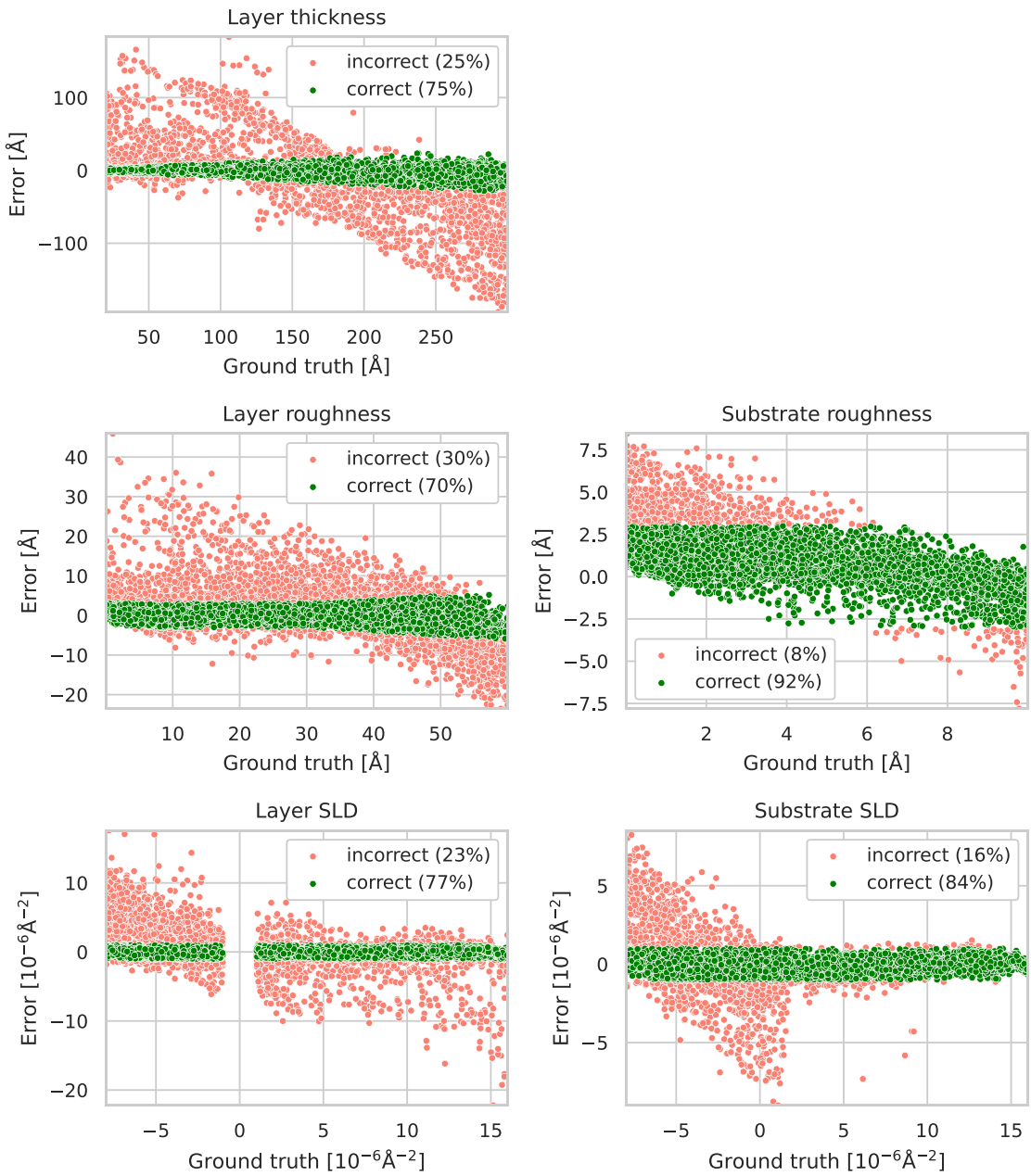

Figure 4. Prediction error of 10 000 simulated reflectivity curves as a function of the ground truth for each of the five thin film parameters. For thicknesses and roughnesses, errors below 3 Å and 10% are considered correct. For SLDs, absolute errors below 1×10^{-6} Å^{-2} are considered correct. Correct and incorrect predictions are colored green and red, respectively.

epoch. This means that the network sees 'fresh' curves every time and is thus forced to generalize more. Since epoch 147 showed the lowest validation loss, we chose the neural network model corresponding to that epoch for all further testing.

3. Results and discussion

3.1. Definition of the prediction accuracy

The performance of the trained neural network was tested using 10 000 simulated reflectivity curves that were generated within the same ranges as the training data, excluding cases with low SLD contrast. The prediction error as a function of the ground truth for each of the five thin film parameters is shown in figure 4. In order to better quantify the performance of the neural network predictions, we separated all predictions into two classes: those that are near the ground truth were classified as 'correct' whereas all others were classified as 'incorrect'. For this, we defined a condition for each parameter under which we consider a prediction 'correct'. For the thickness and roughnesses, all errors smaller than 10% of the ground truth or smaller than 3 Å were considered 'correct'. The absolute condition was added to avoid divergence for small

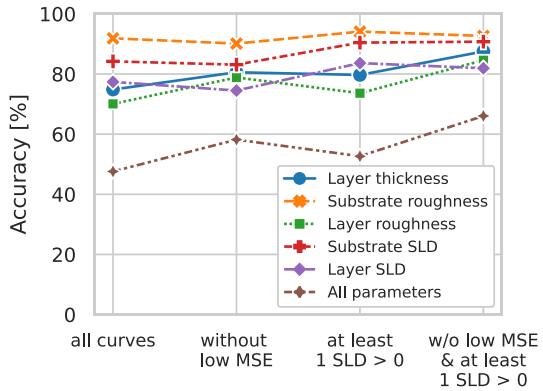

Figure 5. Prediction accuracy of the neural network for the entire test set as well as three smaller subsets that were particularly difficult edge cases have been removed: (1) Removed curves that are incorrect, but still have a very low MSE, (2) Removed curves where layer and substrate have a negative SLD (i.e. no TRE), (3) removed both cases of (1) and (2). By removing difficult edge cases, the accuracy of all parameters being correct is increased from 48% to 66%.

ground truth values. For the SLDs, all errors with an absolute value smaller than 1×10^{-6} Å^{-2} were classified as 'correct'. In figure 4, all 'correct' predictions are colored green whereas all 'incorrect' predictions are colored red. In the following analysis, the percentage of correctly classified predictions out of all predictions will be used to discuss the performance of the neural network model under different circumstances. From here on, we will call this metric the *prediction accuracy*. The distribution of errors for each parameter is given in the supporting information in figure S2.

The prediction accuracy for each parameter of the 10 000 simulated curves is shown in figure 5. For entire test set (i.e. the full parameter space), the accuracy of the individual parameters lies between 71% and 92% whereas the accuracy of all five parameters being correct at the same time is 48%. We consider the latter to be the most important metric for the neural network performance since it represents the likelihood that the neural network predicts all five parameters of an unknown curve correctly. Thus, in the following we will mainly discuss this metric and compare it for different subsets of reflectivity curves in the test set.

3.2. Comparison of prediction accuracy with curve MSE

For conventional curve fitting tools, often the MSE (or an equivalent metric, e.g. chi-squared) between the data and the fitted curve is used to judge the goodness of the fit. This is then used to find a solution to the inverse problem of inferring the correct thin film parameters. However, in some cases, due to a very flat MSE surface with regards to the fitted parameters, there exist many, equally well-fitting curves with different (and potentially wrong) fit parameters. In these cases, reliably extracting the correct thin film parameters is difficult or even impossible without any prior physical knowledge.

To identify especially difficult cases for MSE fitting, we can calculate the MSE

$$E_{\mathrm{MSE}} = \frac{1}{N}\sum_{i}^{N}\left(\log\left(R_i^{\mathrm{gt}}\right) - \log\left(R_i^{\mathrm{p}}\right)\right)^2 \tag{8}$$

between the ground truth curve R^{gt} and the curve simulated via the neural network prediction R^{p} and compare it to the prediction accuracy. In this case, an MSE of 0.1 or lower can be considered an adequate fit whereas an MSE of 0.01 or lower can be considered a near perfect fit. In terms of our testing set, 73% of the predicted curves have an MSE of 0.1 or lower and 46% of 0.01 or lower. When comparing these predictions to the ground truth, we notice that only 60% of the curves with an MSE < 0.01 were actually predicted correctly according to our criteria. Thus, about a fourth of the predictions consist of wrong, but extremely well-fitting neural network predictions that look convincing to a visual inspection. Interestingly, most of these curves are also completely monotonic, which means that these are curves without strong Kiessig oscillations. Strong Kiessig oscillations would ensure that the MSE surface is not flat, therefore reducing the solution space drastically. Prominent reasons for a lack of strong features are a low SLD contrast, very low film thickness or a high roughness compared to the thickness. However, there exist also other conditions that are harder to formalize and attempts at quantifying the information content in reflectivity data using information theory exist [31, 32].

Figure 6(a) shows a curve from the test set with the corresponding prediction from the neural network. The SLD profile of each case is shown in figure 6(b). Although the MSE between the two curves has a very low value of 0.004, the SLD profiles deviate significantly from each other, especially regarding the thickness.

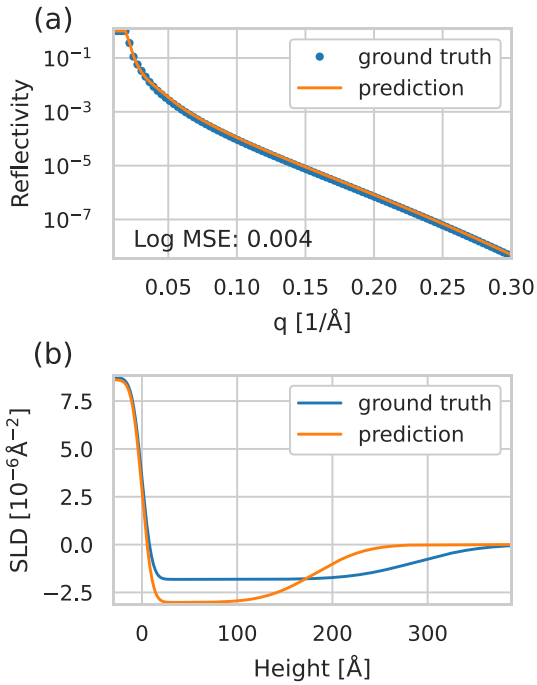

Figure 6. (a) Example of a reflectivity curve and its corresponding neural network prediction where the MSE is especially low although it is classified as incorrect. (b) SLD profiles of the ground truth and the prediction. Although the curves fit perfectly, the SLD profiles differ significantly.

Since there are no visible Kiessig oscillations despite the high thickness of the film, it is likely that many parameter combinations could produce a good fit.

Thus, we conclude that all curves that have a low MSE but are classified as incorrect most likely fall into one of two categories: (1) The fit is actually close to the solution, but falls just outside our accuracy margin. In these cases, the solution is likely already near the MSE minimum and can be reached quickly via a simple gradient descent refinement using the prediction as starting values. (2) The MSE surface is very flat with regards to one or more parameters, leading to multiple solutions with a similar MSE even for large deviations from the ground truth. In these cases, the problem of fitting the data stems from the ambiguity of the data itself and hence it cannot reasonably be expected that the neural network (or another algorithm) can reliably find the true solution.

Therefore, we chose to omit these curves for all following accuracy calculations because they are either probably close enough for refinement or not expected to be feasibly solvable without prior knowledge. After removing those curves, the total accuracy on the test set (all five parameters correct) increases from 44% to 58%.

3.3. Influence of different SLD combinations on the prediction accuracy

An essential part of reflectivity data is the presence of a TRE or a lack thereof. The TRE is located at the critical angle and related to the real part of the SLD contrast $\Delta\rho$ [10] via

$$q_c = \sqrt{16\pi\Delta\rho}. \tag{9}$$

For thicker layers, q_c is given by the SLD contrast between the ambient medium and the layer. For thinner layers, q_c is mainly affected by the contrast between the ambient medium and the substrate. Thus, the TRE contains direct information about the absolute SLDs in the system. Furthermore, it gives a clear way of calibrating measured data, since it is expected that below the TRE almost 100% of intensity is reflected (not accounting for absorption).

To understand the effect of the TRE on the neural network performance, we separated our testing set into four different categories: (1) Both the substrate and layer SLDs are positive, (2) only the substrate SLD is negative, (3) only the layer SLD is negative, and (4) both SLDs are negative. In cases (1) and (3), a TRE is expected to be present in the data since q_c is positive for at least one SLD. Conversely, in the last case there can be no TRE since q_c is negative. In case (2), the TRE depends on the layer SLD but typically, a sharp TRE only forms for higher thicknesses.

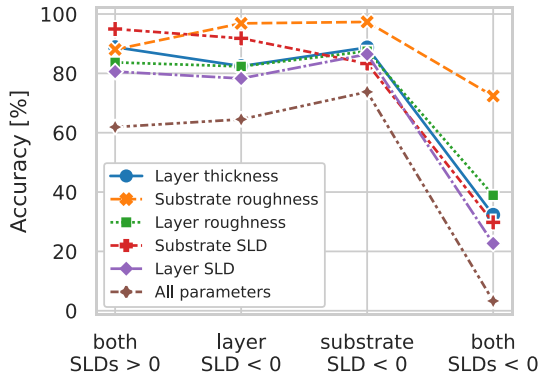

Figure 7. Prediction accuracy for each of the five parameters as well as all five parameters together for each of the four investigated SLD combinations. The accuracy of all parameters drops significantly when both the layer and the substrate SLD are below 0.

The accuracy on the test set for each parameter as well as for all parameters together for each of the four cases is shown in figure 7. It is immediately apparent that the accuracy drops drastically when both the layer and the substrate SLDs are below 0, while for the other three cases it stays in a similar range. When both SLDs are negative, the percentage of predictions where all parameters are correct drops as low as 3%. However, in contrast, for the rest of the cases the accuracy is between 62% and 74%, which is above the average of 58% described in the last section.

This shows that the neural network seems to struggle specifically with cases where there is guaranteed to be no TRE in the data. This hypothesis is supported by the fact that in figure 4, the error of the substrate SLD drastically increases for values smaller than 2×10^{-6} Å^{-2}, which corresponds exactly to a TRE at $q_c = 0.01$ Å^{-1} which in turn is the lowest q value used in our study. Thus, the performance seems to be negatively impacted as soon as the position of the TRE moves below our detectable q range.

From this, we conclude that during the training process, the neural network model 'learned' to extract crucial information from the TRE, so the prediction performance is adversely affected if this information is not available. This is not necessarily unexpected, since the TRE is also an important feature for conventional data analysis. However, note that is that not only the prediction accuracy of the SLD itself is adversely affected, but also all other parameters as well. This might be the result of our chosen neural network design where all parameters are predicted together instead of each parameter being determined independently.

Of course it would be desirable to increase the prediction accuracy for curves without a TRE. While ML models such as neural networks can potentially extract information and make inferences from measured data more efficiently, it is important to bear in mind that these methods are not able to restore missing information. Thus, data with less information encoded in it will always result in lower prediction accuracies. In order to extract the maximum information possible from difficult measurements, in might be useful to train models that are specialized toward specific, difficult edge cases. Since this is beyond the scope of the present study, we removed these low-performing curves (both SLDs < 0) from the test set when discussing the influence of different noise sources on the prediction accuracy in the following section.

3.4. Influence of noise and background on the prediction accuracy

Every real reflectivity measurement contains a number of imperfections such as (among others) noise, background, the angular and energy resolution, the beam profile, the slit settings, the beam divergence and the beam footprint. Thus, training ML models on simulated data without any of those imperfections is likely going to lead to overfitting to features that are not present in real data. Out of these imperfections, this section focuses on the four different sources of noise and background described in section 2.1 and discusses their effect on the prediction accuracy.

The data modifications contained five different background levels $b = \{0, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}\}$, five different Poisson noise levels (equivalent to incident intensity) $s = \{\text{off}, 10^6, 5 \times 10^6, 10^7, 5 \times 10^7\}$, two uniform noise levels (on/off) and two scaling levels (on/off), resulting in a total of 100 different combinations. Each type of modification is described in section 2.1. For each of the combinations, modified variants of the original 10 000 test curves were created and the subsets of difficult edge cases described in the previous sections were removed. For the remaining curves, if no modification was applied, the percentage of completely correctly predicted curves reached 66%, which was the maximum accuracy.

The dependence of the prediction accuracy on each of the modifications is summarized in detail in figure 8 (plots for individual parameters can be found in figures S3–S7. In the top plot, each point refers to

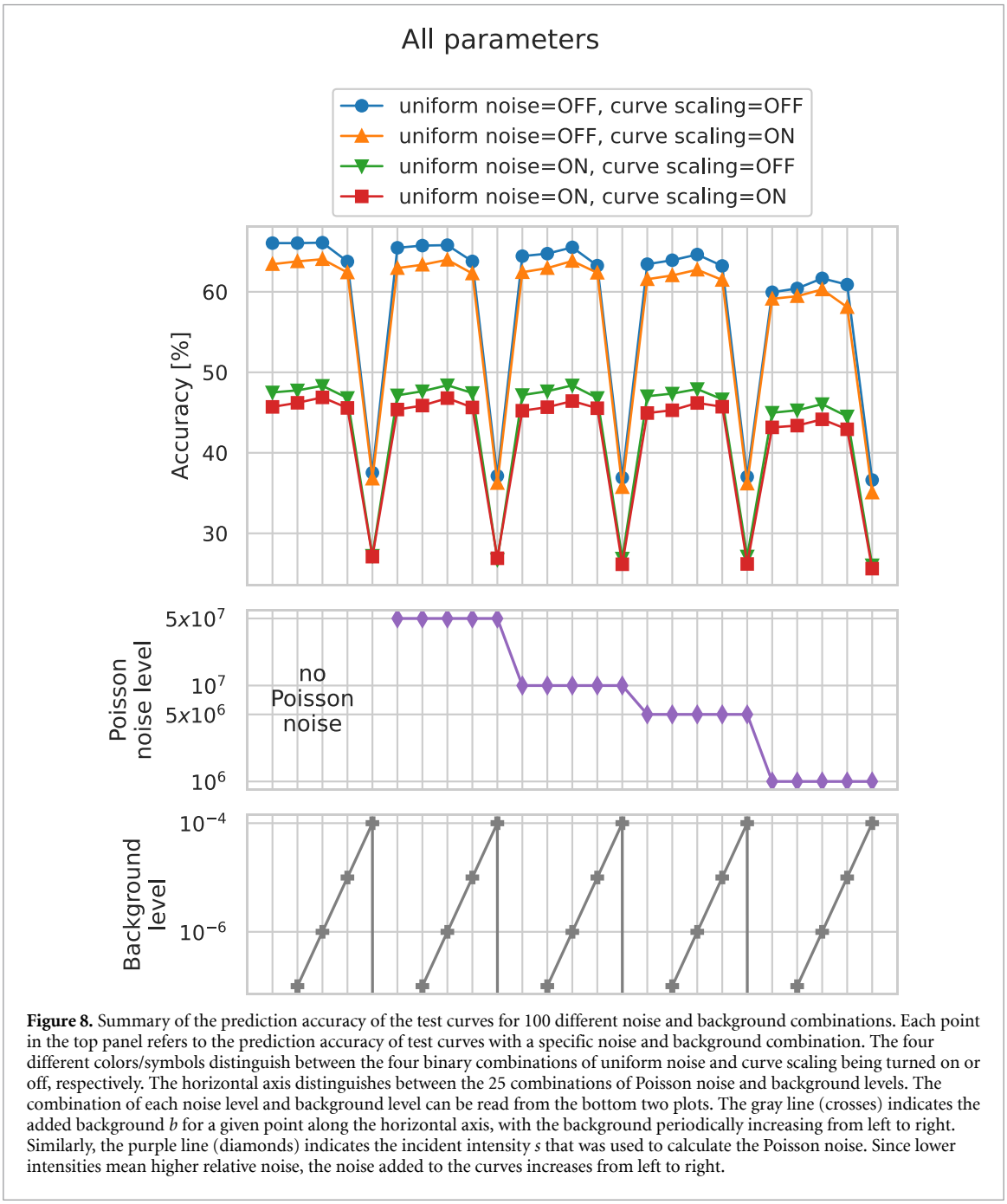

Figure 8. Summary of the prediction accuracy of the test curves for 100 different noise and background combinations. Each point in the top panel refers to the prediction accuracy of test curves with a specific noise and background combination. The four different colors/symbols distinguish between the four binary combinations of uniform noise and curve scaling being turned on or off, respectively. The horizontal axis distinguishes between the 25 combinations of Poisson noise and background levels. The combination of each noise level and background level can be read from the bottom two plots. The gray line (crosses) indicates the added background b for a given point along the horizontal axis, with the background periodically increasing from left to right. Similarly, the purple line (diamonds) indicates the incident intensity s that was used to calculate the Poisson noise. Since lower intensities mean higher relative noise, the noise added to the curves increases from left to right.

the prediction accuracy of a subset of test curves with a specific noise and background combination. The four different colors/symbols distinguish between the four combinations of uniform noise and curve scaling being turned on or off, respectively. The horizontal axis distinguishes between the 25 combinations of Poisson noise and background levels. The combination of each noise level and background level can be read from the bottom two plots. The gray line (crosses) indicates the added background b for a given point along the horizontal axis, with the background periodically increasing from left to right. Similarly, the purple line (diamonds) indicates the incident intensity s that was used to calculate the Poisson noise. Since lower intensities mean higher relative noise, the noise added to the curves increases from left to right.

It is apparent that uniform noise and added background have the strongest impact on the prediction accuracy, whereas curve scaling and Poisson noise only have a minor influence. When curve scaling is turned on, the accuracy is decreased by 1–2 percentage points independent of any other modification. In contrast, Poisson noise seems to only play a role for incident intensities of 10^7 or lower, with its strongest effect at 10^6 where the accuracy is reduced by about 5 percentage points compared to the case without Poisson noise.

When adding background to the curves, the performance is almost not affected at all for background levels up to 10^{-6}. However, for levels of 10^{-5}, we observe a small decrease in accuracy and for 10^{-4}, the

accuracy is reduced by 30 percentage points compared to the unmodified curves. This strongly suggests that there is a critical value above which an additive constant background will obfuscate too much information and therefore make a lot of the curves unsolvable for the neural network.

Furthermore, we note that the accuracy drops by about 20 percentage points when uniform noise is turned on. The difference between Poisson noise and uniform noise is that the latter also affects regions of high intensity, such as the TRE. Thus, the reason why the detrimental effect of uniform noise is relatively strong might be related to information that is encoded in the TRE as described in the last section. By modifying the TRE, some of the information might be lost. The same is likely true for the curve scaling, since it also affects the TRE, but the scaling factor might not be large enough to produce a strong effect. Interestingly, the decrease in accuracy caused by curve scaling and uniform noise compounds with the effect of a high background. This suggests that these effects are independent from each other, since a high background mainly affects high-q features while scaling and uniform noise affect mainly low-q features (e.g. the TRE).

3.5. Other challenging cases and prospects

We note that in addition to the issues discussed here, further challenges may arise from more complex situations given by the systems under study. First of all, we have assumed a box-like SLD. If the SLD exhibits a profile incompatible with the approximation by a box, e.g. a sloping or graded profile, further and more elaborate training is most likely needed. Second, samples that consist of multiple layers (i.e. beyond one layer on a substrate) are obviously not yet included in our study. The incorporation is in principle straight-forward, but of course the solution space increases with the number of parameters and difficulties are expected when different combinations or orders of layers result in similar XRR or NR curves. In these cases it might be necessary to constrain the solution space of the inverse problem by providing any available *a priori* knowledge about the studied system to the neural network.

For some applications, it is common to combine multiple data sets during analysis. This is certainly possible for ML approaches, but of course requires a broader training strategy. One application may be reflectivity time series during growth, annealing or oxidation experiments, where it is beneficial to fit all XRR or NR curves of a series together. By applying boundary conditions, such as demanding a monotonically increasing thickness, the ambiguities in the analysis can potentially be reduced. Another example concerns NR from magnetic structures where several data sets with different polarization ($\uparrow\uparrow$, $\uparrow\downarrow$, etc as well as spin flip and non-spin flip) need to be fitted simultaneously, possibly with prior knowledge from XRR measurements to determine the chemical structure. Again for this situation, as well as for possibly others, there is no fundamental reason why ML could not be employed given a sufficiently large and varied training data set.

4. Conclusion

The results discussed in this work provide insights into the behavior of neural networks when predicting thin film parameters from reflectivity data. These insights are necessary to understand which types of reflectivity can be processed easily by the neural network and which are more difficult. This understanding will help to further improve and adapt the design of ML models to the specific needs of scattering data for which a simple inversion is not possible. The results show that three subsets of reflectivity data seem to be particularly difficult for the neural network: (1) Curves with ambiguous solutions where the MSE surface between the curve and the fit as a function of the parameters is flat, (2) curves where the SLD of both the layer and the substrate are negative (i.e. no TRE), and (3) curves with noise on low-q features or particularly high background.

When tested on noise-free data, the trained neural network was able to predict all five thin film parameters 48% of the time (individual accuracies ranged from 71% to 92%). We identified that subsets (1) and (2) mainly consist of curves that lack information-rich features such as oscillations or a TRE. Thus, by removing these cases from the test set the accuracy increased to 66% (82%–93% for individual parameters).

By further studying the influence of different noise and background sources, we showed that, if included in the training set, most curve modifications do not significantly impact the prediction accuracy. However, if a critical threshold of 10^{-4} for the background was crossed, the accuracy dropped significantly. Fortunately, this value is rarely exceeded in experimental data, since high backgrounds are usually already subtracted at the data reduction step, leaving only a smaller residual background. Thus, background is not likely to play an important role when using the neural network on real data.

Furthermore, when moderate uniform noise was applied to the data, the performance of the neural network was noticeably affected. From our analysis, we concluded that this was most likely due to its effect on the TRE. This further shows the importance of the TRE as a source of information and together with the SLD dependence of the performance indicates that the neural network model has 'learned' to extract critical

information from the TRE. Moreover, this suggests that the model is not overfitting to the simulation, but is instead picking up on real physical features of the data.

To further improve the neural network performance in the future there are two obvious strategies, both of which rely on a narrowing of the task. Firstly, one might choose to narrow the task of the neural network to cases with clear solutions and remove classes of difficult edge cases from the training set. This approach might be favorable when the experimental data is expected to have clear features where training with data without clear features would only serve to make the task harder without any benefit. A second approach might be to select a subset of difficult edge cases that are most likely to appear in the experimental data (such as curves without a TRE) and create a training set that focuses on these cases.

The present work may also give an indication about the information content of different reflectivity curves. Under the premise that neural network is able to extract close to the maximum amount of information from a reflectivity curve, the difficult curves identified herein may also be curves with a low amount of physical information. Therefore, simulations and predictions with our neural network may help with experimental design through identifying ambiguous and difficult measurement results and then avoiding these parameter combinations or complementing them with additional information.

Since the results of this work are based on simulated data, future efforts should also be focused on translating these achievements to experimental data. To this end, it is necessary to investigate other data imperfections such as a finite angular and energy resolution and the influence of the beam shape and their effect on the prediction accuracy.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This research was part of a project funded by the German Federal Ministry for Science and Education (BMBF). Frank Schreiber is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

ORCID iDs

Alessandro Greco https://orcid.org/0000-0002-3714-7941
Vladimir Starostin https://orcid.org/0000-0003-4533-6256
Alexander Hinderhofer https://orcid.org/0000-0001-8152-6386
Alexander Gerlach https://orcid.org/0000-0003-1787-1868
Maximilian W A Skoda https://orcid.org/0000-0003-0086-2965
Stefan Kowarik https://orcid.org/0000-0001-7661-1949
Frank Schreiber https://orcid.org/0000-0003-3659-6718

References

[1] Tolan M 1999 *X-ray Scattering from Soft-Matter Thin Films: Materials Science and Basic Research* (*Springer Tracts in Modern Physics*) (Berlin: Springer)
[2] Holý V, Pietsch U and Baumbach T 1999 *High-Resolution X-Ray Scattering from Thin Films and Multilayers* (*Springer Tracts in Modern Physics*) (Berlin: Springer)
[3] Braslau A, Pershan P S, Swislow G, Ocko B M and Als-Nielsen J 1988 Capillary waves on the surface of simple liquids measured by x-ray reflectivity *Phys. Rev.* A **38** 2457
[4] Neville F, Cahuzac M, Konovalov O, Ishitsuka Y, Lee K Y C, Kuzmenko I, Kale G M and Gidalevitz D 2006 Lipid headgroup discrimination by antimicrobial peptide ll-37: insight into mechanism of action *Biophys. J.* **90** 1275
[5] Skoda M W A, Thomas B, Hagreen M, Sebastiani F and Pfrang C 2017 Simultaneous neutron reflectometry and infrared reflection absorption spectroscopy (IRRAS) study of mixed monolayer reactions at the air–water interface *RSC Adv.* **7** 34208
[6] Russell T P 1990 X-ray and neutron reflectivity for the investigation of polymers *Mater. Sci. Rep.* **5** 171
[7] Lehmkühler F, Paulus M, Sternemann C, Lietz D, Venturini F, Gutt C and Tolan M 2009 The carbon dioxide-water interface at conditions of gas hydrate formation *J. Am. Chem. Soc.* **131** 585
[8] Kowarik S, Gerlach A, Sellner S, Schreiber F, Cavalcanti L and Konovalov O 2006 Real-time observation of structural and orientational transitions during growth of organic thin films *Phys. Rev. Lett.* **96** 125504
[9] Skoda M W 2019 Recent developments in the application of x-ray and neutron reflectivity to soft-matter systems *Curr. Opin. Colloid Interface Sci.* **42** 41
[10] Sivia D S 2011 *Elementary Scattering Theory: For X-ray and Neutron Users* (Oxford: Oxford University Press)
[11] Daillant J and Gibaud A 2009 *X-Ray and Neutron Reflectivity* (Berlin: Springer)

[12] Hoogerheide D P, Heinrich F, Maranville B B and Majkrzak C F 2020 Accurate background correction in neutron reflectometry studies of soft condensed matter films in contact with fluid reservoirs *J. Appl. Crystallogr.* **53** 15
[13] Parratt L G 1954 Surface studies of solids by total reflection of x-rays *Phys. Rev.* **95** 359
[14] Abelès F 1950 La théorie générale des couches minces *J. Phys. Radium* **11** 307
[15] Heavens O S 1955 *Optical Properties of Thin Solid Films* (London: Butterworths Scientific Publications)
[16] Gerelli Y 2016 *Aurore*: new software for neutron reflectivity data analysis *J. Appl. Crystallogr.* **49** 330
[17] Kienzle P, Krycka J, Patel N and Sahin I 2011 Refl1d [computer software] (available at: http://reflectometry.org/danse)
[18] Nelson A R J 2006 Co-refinement of multiple-contrast neutron/x-ray reflectivity data using motofit *J. Appl. Crystallogr.* **39** 273
[19] Nelson A R J and Prescott S W 2019 *refnx*: neutron and x-ray reflectometry analysis in python *J. Appl. Crystallogr.* **52** 193
[20] Björck M and Andersson G 2007 Genx: an extensible x-ray reflectivity refinement program utilizing differential evolution *J. Appl. Crystallogr.* **40** 1174
[21] Danauskas S M, Li D, Meron M, Lin B and Lee K Y C 2008 Stochastic fitting of specular x-ray reflectivity data using *J. Appl. Crystallogr.* **41** 1187
[22] Franke D, Jeffries C M and Svergun D I 2018 Machine learning methods for x-ray scattering data analysis from biomacromolecular solutions *Biophys. J.* **114** 2485
[23] Ikemoto H, Yamamoto K, Touyama H, Yamashita D, Nakamura M and Okuda H 2020 Classification of grazing-incidence small-angle x-ray scattering patterns by convolutional neural network *J. Synchrotron Rad.* **27** 1069
[24] Chang M-C, Wei Y, Chen W-R and Do C 2020 Deep learning-based super-resolution for small-angle neutron scattering data: attempt to accelerate experimental workflow *MRS Commun.* **10** 11
[25] Liu S, Melton C N, Venkatakrishnan S, Pandolfi R J, Freychet G, Kumar D, Tang H, Hexemer A and Ushizima D M 2019 Convolutional neural networks for grazing incidence x-ray scattering patterns: thin film structure identification *MRS Commun.* **9** 586
[26] Park W B, Chung J, Jung J, Sohn K, Singh S P, Pyo M, Shin N and Sohn K-S 2017 Classification of crystal structure using a convolutional neural network *IUCrJ* **4** 486
[27] Vecsei P M, Choo K, Chang J and Neupert T 2019 Neural network based classification of crystal symmetries from x-ray diffraction patterns *Phys. Rev.* B **99** 245120
[28] Greco A, Starostin V, Karapanagiotis C, Hinderhofer A, Gerlach A, Pithan L, Liehr S, Schreiber F and Kowarik S 2019 Fast fitting of reflectivity data of growing thin films using neural networks *J. Appl. Crystallogr.* **52** 1342
[29] Abadi M *et al* 2016 Tensorflow: large-scale machine learning on heterogeneous distributed systems (arXiv:1603.04467) [cs.DC]
[30] Kingma D P and Ba J 2017 Adam: a method for stochastic optimization (arXiv:1412.6980) [cs.LG]
[31] Sivia D and Webster J 1998 The Bayesian approach to reflectivity data *Physica* B **248** 327
[32] Treece B W, Kienzle P A, Hoogerheide D P, Majkrzak C F, Lösche M and Heinrich F 2019 Optimization of reflectometry experiments using information theory *J. Appl. Crystallogr.* **52** 47