Machine learning for neutron scattering at ORNL*

IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 023001 https://doi.org/10.1088/2632-2153/abcf88

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
14 July 2020

REVISED
26 October 2020

ACCEPTED FOR PUBLICATION
1 December 2020

PUBLISHED
29 December 2020

PERSPECTIVE

Machine learning for neutron scattering at ORNL*

Mathieu Doucet[1], Anjana M Samarakoon[1], Changwoo Do[1], William T Heller[1], Richard Archibald[2], D Alan Tennant[3,4], Thomas Proffen[1] and Garrett E Granroth[1]

[1] Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
[2] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States of America
[3] Materials Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831 United States of America
[4] Shull Wollan Center—A Joint Institute for Neutron Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831 United States of America

E-mail: doucetm@ornl.gov

Keywords: neutron scattering, machine learning, spectroscopy, diffraction, sans, super resolution

Abstract

Machine learning (ML) offers exciting new opportunities to extract more information from scattering data. At neutron scattering user facilities, ML has the potential to help accelerate scientific productivity by empowering facility users with insight into their data which has traditionally been supplied by scattering experts. Such support can help in both speeding up common modeling problems for users, as well as help solve harder problems that are normally time consuming and difficult to address with standard methods. This article explores the recent ML work undertaken at Oak Ridge National Laboratory involving neutron scattering data. We cover materials structure modeling for diffuse scattering, powder diffraction, and small-angle scattering. We also discuss how ML can help to model the response of the instrument more precisely, as well as enable quick extraction of information from neutron data. The application of super-resolution techniques to small-angle scattering and peak extraction for diffraction will be discussed.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have recently been the focus of increased attention in both science and the private sector [1]. ML techniques, a subset of AI, are computational algorithms that become better at performing tasks by being trained using example data. Although neural networks, a common ML technique, have been used in scientific data analysis for decades, recently developed packages such as TensorFlow [2], scikit-learn [3], and Keras [4] make ML techniques broadly accessible to the scientific community. While ML opens the door to new possibilities for the interpretation of scientific data, significant effort is still required to identify which problems can best be tackled with ML and how. Recent studies have explored how the US scientific user facilities can leverage ML as a tool to accelerate discovery [5].

ML techniques are particularly useful for efficient extraction of correlations in large and complex data sets. Advances in neutron sources and instrumentation allow increasingly higher data rates and extend the range of possibilities for improved time-resolved and multi-modal studies of materials [6]. New analysis tools are needed to take full advantage of this wealth of data. Furthermore the materials of interest are becoming much more complex. Distinguishing between several structural models for small-angle neutron scattering (SANS), multiple sources of short range order in diffraction, novel quantum magnetic states from each other and complex classical spin states, are all examples of analysis challenges arising from the current complexity of the samples. New tools and methods are needed to quickly work with these complex systems.

The way scientific user facilities help their users plan and execute their experiments is also in need of automation. The same techniques used to analyze large and complex data sets can be used to perform fast

analysis of neutron data as it is acquired. The ability to quickly extract structural information during data acquisition will enable fast feedback systems to help automate measurements using both neutron, sample environment data, *in situ* characterization, and external data from other sources [7].

Oak Ridge National Laboratory (ORNL) hosts two world leading neutron scattering facilities. The Spallation Neutron Source [8] (SNS) and the high flux isotope reactor perform nearly 800 user experiments per year. These two facilities host 30 unique instruments that cover the modalities of neutron scattering science; namely diffraction, spectroscopy, small-angle scattering, reflectometry, and imaging. These instruments produce almost 500 publications per year in a wide range of sciences including biology, chemistry, materials science, and physics. With the question of whether ML can help accelerate discovery at neutron facilities in mind, several scientists at ORNL have been exploring the use of ML in the analysis of neutron scattering data. Though independent efforts, they are grouped into two specific focuses for clarity in presentation: the inverse scattering problem and characterization of the instrument response function.

Before detailing the work on these two problems, a brief discussion of neutron scattering follows. Scattering experiments provide invaluable atomic to micro scale information on structure and dynamics. They directly probe the response function,

$$G(\mathbf{r},t) = \int_0^\infty < \rho(\mathbf{r}_0,0)\,\rho(\mathbf{r}_0+\mathbf{r},t) > \mathrm{d}\mathbf{r}_0 \tag{1}$$

of the material where $\rho(\mathbf{r},t)$ is the density operator as a function of position and time [9][5]. The neutron scattering measurement probes this function through a diffraction process which measures the Fourier transform of the function namely $S(\mathbf{Q},\omega)$, where $\mathbf{Q}$ is momentum and ω is energy. In the case where only the structure is studied, $G(\mathbf{r},\infty)$ is the quantity of interest. It is accessed via its Fourier transform $S(\mathbf{Q})$, the quantity measured on diffractometers for atomic scale structures and SANS instruments for nano scale to microscopic structures. As diffraction processes only measure intensities, phase information is lost in the measurement [10, 11]. Furthermore limited detector coverage and time resolution, experimental background, and instrumental artifacts present fundamental limits to the technique [12]. All of these inhibit the direct computation of a model from the distribution of scattered particles. Such a problem is called the inverse scattering problem and has generally not been solvable. The direct scattering problem, on the other hand, is the computation of the correlation function that quantifies the neutron scattering from a material for a given model. Approximate solutions to these models are usually computed, and a lengthy period of iterative analysis is often required to obtain a good agreement between model and data. Depending on the problem, it can take over a year to interpret neutron scattering experiments, making it the major bottleneck in terms of time, expertise, and effective use of experimental facilities [13].

High performance computing (HPC) and ML algorithms offer new ways to solve this problem. Numerous simulations of scattering outcomes, using validated simulations of materials, can be processed by artificial neural networks to develop comprehensive mappings between underlying materials' characteristics and a scattering measurement. These mappings provide the characteristics to be directly inferred from the scattering, and therefore solve the inverse scattering problem. The constraints in the physical system are accounted for in the simulations and therefore are inherent to the mappings. Among others [14, 15], specific efforts at ORNL have been exploring the use of ML to tackle the inverse problem for several modalities. Work has been done on predicting scattering length density profiles in reflectometry [16], Hamiltonian solution from magnetic diffuse scattering [12], predicting structures from powder data [17], and choosing the appropriate SANS models [18, 19]. The last three cases will be summarized in more detail in sections 2, 3, and 4.1.

For the instrument response function, two cases where ML has assisted in analysis by its characterization are discussed. Section 4.2 describes using ML on fine resolution SANS data to understand the response function [20]. This learned resolution response is used, along with super resolution methods, to obtain finer resolution results from coarse resolution data. Finally, section 5 shows that ML can be used to determine the shape parameters of a Bragg reflection from a large number of strong reflections [21]. This peak shape information is then used for better integration of weak peaks.

2. Machine learning for solving Hamiltonians of magnetic systems

Understanding the origins of the magnetic scattering from samples is a prime example where being able to solve the inverse problem would greatly expedite discovery. Reverse Monte Carlo approaches [22], based on

[5] Neutrons are sensitive also to magnetic correlations via scattering through their dipole moment. This cross-section probes the correlation functions of spin moments $\mathbf{S}$ with components S^α $(\alpha = x, y, z)$ as $\langle S^\alpha(\mathbf{r}_0,0)S^\beta(\mathbf{r}_0+\mathbf{r},t)\rangle$.

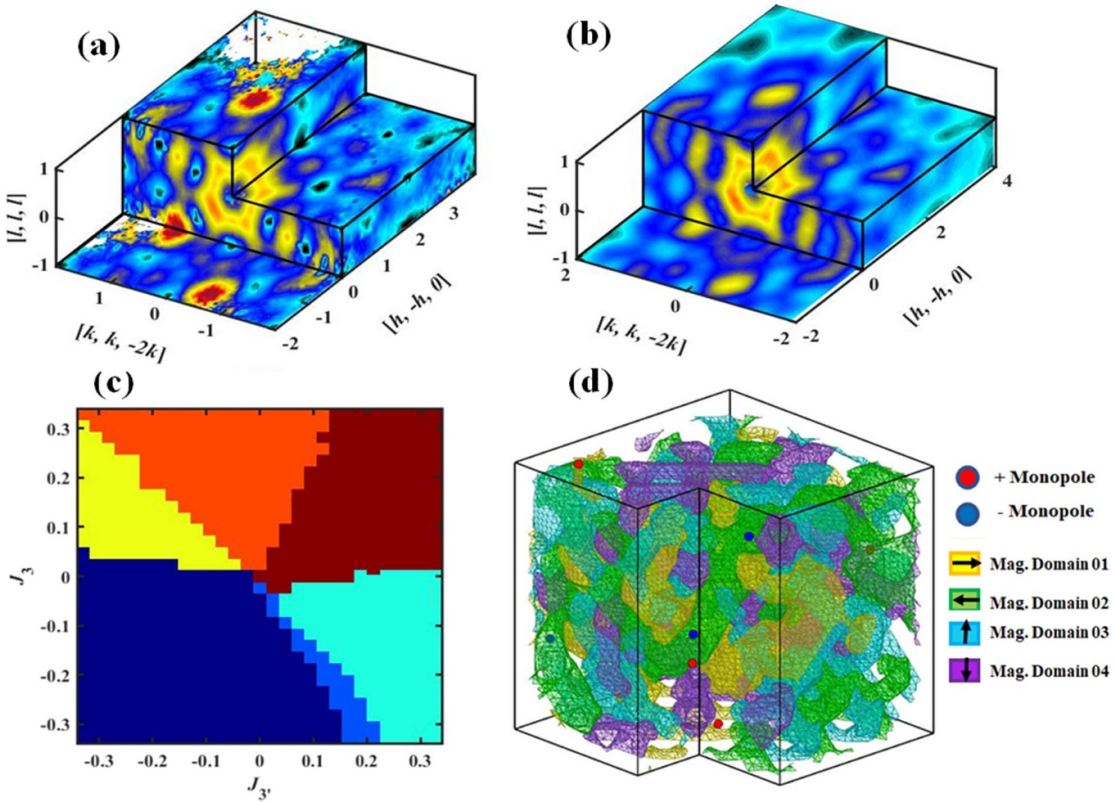

Figure 1. Machine learning applied to $Dy_2Ti_2O_7$: ML using trained autoencoders results in visualization of glass formation in a cooled spin liquid. A large number of simulations of neutron experiments in a high dimensional model space $\{\mathcal{H}(\mathcal{J})\}$ via HPC. Optimization of the model parameters to the measured data is undertaken in the autoencoder latent space which is robust to background and instrumental artifacts. Panel (a) is a rendering of a 3D diffuse scattering measurement of the spin correlations $\mathcal{S}_{\rm exp}(\mathbf{Q})$ in the gauge spin liquid $Dy_2Ti_2O_7$ measured as one of 40 data sets on CORELLI. Panel (b) is the result of denoising and removal of experimental artifacts by the autoencoder ($\mathcal{S}_{AE}(\mathbf{Q})$) which provides optimized model parameters from the 3D data set. (c) Shows a slice through the model space of interaction parameters where ML has identified distinct physical behaviors (phases) and learned features, (d) visualization of the model in real space showing monopole localization and intertwined order formation that constitutes the emergent low temperature glass phase. Details of methodology are in the text and reported in [12]. Versions of figures ((a)–(c)) appeared originally in Samarakoon *et al* 2020 *Nature Communications* **11** 892 [12] and are reproduced under Creative Commons License 4.0. All figures ((a)–(d)) were originally created by A. Samarakoon.

the Metropolis-Hastings algorithm, have been extensively used to find likely spin configurations compatible with diffraction measurements of $\mathcal{S}(\mathbf{Q})$. These invert to real space correlations via entropically-selected representative spin configurations rather than the models themselves. Using ML trained over large numbers of simulations is an alternative strategy to solve the inverse problem to a model instead and has been successfully demonstrated on diffuse scattering in $Dy_2Ti_2O_7$ (DTO) [12]. In this study it was shown to provide direct benefits in terms of providing insight into the physical behavior as well as guiding measurement strategies and handling of experimental artifacts. It is also being applied to inelastic scattering to analyze data, see figure 1. A brief review of the specifics are provided below.

For spin systems the density ρ in equation (1) is replaced by the magnetization density which for localized spins is proportional to S^{α}, where α indicates the directional component of this vector quantity [23, 9]. Time-of-flight techniques (for example, see [24]) yield the scattering functions over large volumes of wave-vector and energy. While some deductions about the magnetic state in a material can be made by interpreting the scattering directly, e.g. an antiferromagnet versus a ferromagnet, this gives little insight into the stability of the magnetic phase or what is driving its formation. The energetics of the underlying system is what determines this and is thus critically important. The central problem then in understanding magnetic systems is to extract the Hamiltonian $\mathcal{H}$, which is the model of the interacting system, and modeling, understanding, and exploring its behavior.

The Hamiltonian is usually characterized by a few interaction parameters $\{\mathcal{J}\}$. However in quantum materials there are additional interactions from spin–orbit coupling leading to complex models with multiple parameters. The main bottleneck for understanding these materials is the difficulty of extracting the Hamiltonian and its parameters in a reliable way.

Essential to undertaking ML approaches are methods for calculating $\mathcal{S}^{\alpha\beta}(\mathbf{Q},\omega)$ from a given Hamiltonian. Linear spin-wave theory and mean field approaches are fast, their codes are readily

available [25, 26], and they are useful in many applications. Of relevance to the current discussion, they have been shown to be useful in using ML to find classical spin wave solutions [27]. Monte Carlo and Landau Lifshitz dynamics, however, are more versatile for semi-classical modeling of magnetic materials [28] because they are capable of treating non-linear physical behaviors, which are essential for many complex magnetic materials, as well as thermal effects. In addition, they provide a computational basis for the calculation of other commonly measured physical properties such as heat capacity and magnetic susceptibility. This latter approach has therefore been used in this study. In order to perform the calculations on the scale needed for ML, HPC codes have been developed that calculate $\mathcal{S}^{\text{sim}}(\mathbf{Q},\omega)$ given $\mathcal{H}\{\mathcal{J}\}$ [29, 30] in order to leverage the Oak Ridge Leadership Computing Facilities [31].

The actual intensity measured on the spectrometer $\mathcal{S}^{\text{exp}}(\mathbf{Q},\omega)$ is convoluted with the instrumental resolution, modulated by scattering correction factors, and combined with undesired scattering from phonons, sample mountings, sample environment, *etc*. In DTO the data collected were of diffuse scattering, measured on Corelli [32], where the measurement process integrates over energy, yielding an $\mathcal{S}^{\text{exp}}(\mathbf{Q})$ which can be compared to the equal-time correlations $\mathcal{S}(\mathbf{Q})$.

2.1. Dimensionality reduction by an autoencoder neural network

Now that the connection between the experiment and the model calculations has been summarized, the application of ML to understand $\mathcal{H}$ is described. A first approach to this problem would be least squares minimization, but noise in the simulations and the aforementioned artifacts in $\mathcal{S}^{\text{exp}}(\mathbf{Q})$ make conventional least squares minimization unreliable. Instead, ML has been used to automate the data analysis. To begin, the allowed symmetries and rough strengths of interactions were estimated to provide a domain of values and Hamiltonians $\{\mathcal{H}(\{\mathcal{J}\})\}$. This domain, covered by five to ten independent parameters, provides a wide variety of magnetic moment configurations and dynamics which are reflected in the spin correlations. In DTO, $\mathcal{J}$ comprises four independent magnetic exchange interactions as well as dipolar magnetic coupling. Next, an autoencoder neural network is trained to learn a compressed representation of simulated three-dimensional scattering data ($\mathcal{S}_{\text{sim}}(\mathbf{Q})$), over the wide domain of interactions $\mathcal{H}\{\mathcal{J}\}$ calculated above, and at finite temperatures matching experimental conditions sufficiently broad to cover all potentially important characteristic features of the DTO scattering data. Specifically, the autoencoder takes as input $\mathcal{S}_{\text{sim}}(\mathbf{Q})$, encodes it into a compressed latent space representation S_L, and then decodes it to an output $\mathcal{S}_{AE}(\mathbf{Q})$. This output captures the essence of the input $\mathcal{S}_{\text{sim}}(\mathbf{Q})$ while removing irrelevant noise and artifacts.

Next the optimization of the model $\mathcal{H}\{\mathcal{J}\}$ through random sampling of the Hamiltonians is iterative. An autoencoder-based error measure, $\chi^2_{S_L} = \sum_L \left(S^{\text{exp}}(L) - S^{\text{sim}}(L)\right)^2$, where $S(L)$ is the latent space representation of $S(Q)$, is found to be more sensitive to features in the data and less biased by artifacts [12] than the conventional least squares between data and model. Indeed $\chi^2_{S_L}$ is more robust to stochastic noise and allows more precise estimation of $\mathcal{J}$. For each iteration, random Hamiltonians are sampled and all available data at each interaction is used to build $\hat{\chi}^2_{S_L}$, the low-cost estimator for $\chi^2_{S_L}$. Randomly selected Hamiltonians are included in the dataset, each being sampled uniformly, subject to the constraint $\hat{\chi}^2(\mathcal{H}) < c$. The cut-off parameter c decreases exponentially, rescaling by a factor 0.9 at each iteration. Consequently, later iterations in the optimization procedure are focused on regions where $\chi^2_{S_L}$ is smallest. The optimization procedure terminates after about 40 iterations, at which point $\mathcal{H}_{\text{best}}$ is taken as the minimizer of $\hat{\chi}^2_{S_L}(H)$. Furthermore since the Landau Lifshitz code also calculates magnetic susceptibility, and heat capacity, experimental data from these experiments was also used to constrain the minimization providing even finer confidence intervals for models [12]. Validation tests indicate that the inferred Hamiltonian $\mathcal{H}_{\text{best}}$ accurately predicts temperature and field dependence of both the structure factor and magnetization as well as the glass formation and irreversibility characteristics of out-of-equilibrium behavior. In summary, ML, on simulated neutron scattering and heat capacity data, has been used to provide a better determination of the Hamiltonian for DTO by comparison to the experimental data.

2.2. Classification by a clustering algorithm

Other ML techniques can be used on the large number of simulated data sets to provide a greater understanding of the system. Specifically agglomerative hierarchical clustering algorithms were used [33] to classify and map out the regions of different features as a function of $\{\mathcal{J}\}$. These were applied to the same simulations used to train the autoencoder and their corresponding $\mathcal{S}^{\text{sim}}(\mathbf{Q})$. The clustering algorithm requires as input the pairwise distances between all points in the data-set and the squared distance in the autoencoder latent space was used. $\mathcal{S}(\mathbf{Q})$ and $\mathcal{S}(\mathbf{Q},\omega)$ are particularly information rich physical observables, so they are excellent discriminators for classes of physical behaviors and phases. As such the auto classification is of great importance to the physical understanding of the system and can be used for both ordered and disordered materials. Access to such mappings aids in forming materials modification strategies and the identification of which external parameters, such as magnetic field, temperature, and pressure, need

to be applied to stabilize new phases and reveal new physics. They can also be used to plan new experiments. The large dimensional space involved in $\{\mathcal{H}(\mathcal{J})\}$ necessitates the use of ML approaches and represents a new capability for researchers.

In summary, the two ML approaches of sections 2.1 and 2.2 provided direct insight into the physical behavior of DTO [12]. In identifying discriminating features the autoencoder abstracts out the essential aspects of the correlations responsible for stability of phases and driving phase transitions; thus providing important new understanding.

A few general points from this work should be emphasized. First as it focuses on essential features in the data the autoencoder eliminates background noise and artifacts in raw scattering data. Second the methodology is general. The magnetic Hamiltonian codes could be replaced by other codes that generate $G(r,t)$. This makes it readily applicable to other materials and scattering problems including inelastic scattering and non-magnetic systems. Of particular interest for magnetic system are methods with higher levels of quantum entanglement such as Schwinger bosons, density matrix renormalization group, and exact diagonalization. These all require significantly more computational time and resources, so HPC is essential. Hybrid approaches with these codes and the semiclassical Monte Carlo and Landau Lifshitz dynamics methods would provide an even more powerful methodology. Continued efforts by Samarakoon and coworkers are showing that these approaches succeed when applied to complex quantum systems and that they are extensible to dynamical systems. These results are in preparation for publication.

Finally, the ability of ML to work in 3D, in the case of diffuse scattering, and 4D for single crystal inelastic scattering, is significant. A longstanding barrier to working with the large data sets produced by neutron time-of-flight techniques has been the human constraints on understanding high dimensional data.

3. Material structure prediction

The ability to determine the atomic structure from diffraction data has revolutionized our understanding of materials and their properties over the past decades. However, structure solution and refinement is still a largely manual process using direct problem methods and relying on the prowess of the individual researcher. Machine learning offers a promising approach to address this problem and solve structures. Liu and coworkers recently published an approach to determine the space group of a structure from the atomic pair distribution function (PDF) using ML [34]. The PDF is the powder averaged $G(< \mathrm{r} >, \infty)$, derived from the measured diffraction data via a Fourier transform (see e.g. [35]). Their approach used ML models trained on more than 100 000 PDFs calculated from structures in the 45 most heavily represented space groups. The trained a convolutional neural network (CNN) identified the correct space group among the top six most likely groups 91.9% of the time.

While the study by Liu [34] used the PDF as data, Garcia-Cardona and coworkers used powder diffraction data directly to train the network [17]. Their workflow is depicted on the right of figure 2. The neural network provides a structural model prediction using the calculated powder diffraction data, accelerating discovery and aiding experts in structure solution and refinement. With ever increasing data collection rates and sample synthesis capabilities, this approach has the potential to alleviate the bottleneck created by the need for scientist input and speed up scientific discovery.

The target structures for the training set used in [17] were perovskite with the general formula ABO_3. The goal was to demonstrate the ability to predict space group and structural parameters such as lattice parameters and atomic displacement parameters accurately enough to enable a successful Rietveld refinement [36] from the derived model. Training diffraction data were calculated using the GSAS-II software [37] and more than 1 500 000 data sets were generated.

The ML framework depicted on figure 3 has two components: (a) a classifier to predict the crystallographic class and (b) a regressor to predict structural parameters such as lattice parameters.

To understand the relationship between the several crystallographic classes Garcia-Cardona and coworkers [17] applied multidimensional scaling to random subsets of 500 examples from each class. Their analysis showed that while simple distance-based learning algorithms (such as k-nearest neighbors) would easily distinguish the cubic, tetragonal, and trigonal classes, there was considerable overlap between the monoclinic and triclinic classes, indicating that any similar distance-based algorithms will likely be unable to differentiate the two classes. As a next step, CNN was developed, applying 1D convolutions directly to the diffraction data $I(Q)$ for classification. This CNN-based classifier differentiates the cubic, tetragonal, and trigonal classes with near 100% accuracy, only classifying three trigonal examples out of 3000 total incorrectly. On the triclinic and monoclinic classes the CNN-based classifier made successful predictions with about 92.65% accuracy, leaving considerable room for improvement.

After classification, a regressor was used to predict the unit cell parameters corresponding to the particular predicted symmetry class. The authors found that using Random Forest regression gives good

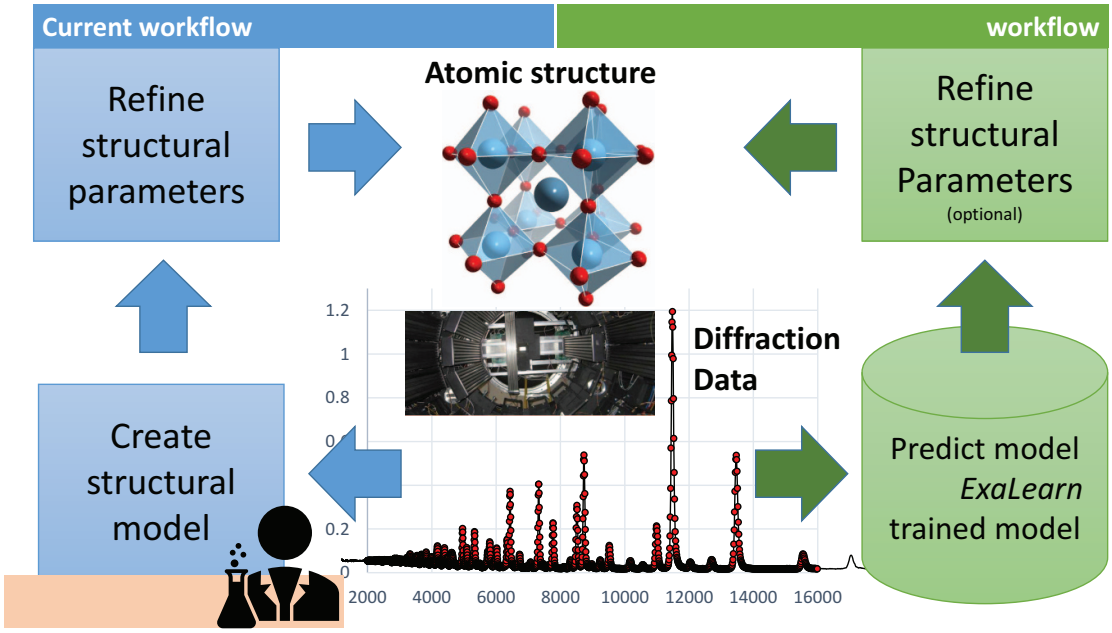

Figure 2. Typical workflow in structure determination and refinement. The loop on the left show the current workflow involving a domain expert creating a structural model. The loop on the right is the workflow described here where the structural model and parameter estimated are obtained from a trained network [17].

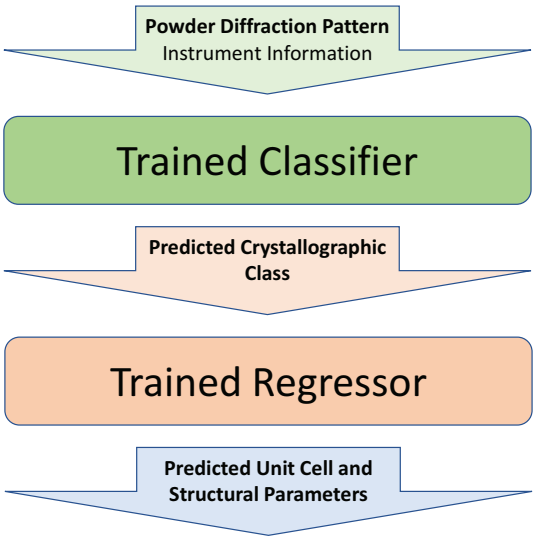

Figure 3. The ML pipeline consists of a classifier to identify the crystallographic class and a regressor predicts the structural parameters [17].

predictions for lattice parameters when using simulated test data. However, when using actual experimental data measured on NOMAD [38] at SNS, only the cubic case gave reasonable results.

While this prototype clearly shows the potential of the structural solution from diffraction data using ML, it also highlights the many obstacles that need to be overcome. Many ML approaches in neutron scattering rely on synthetic or simulated data for training while the resulting network is then used on experimental data. This approach brings up many interesting questions regarding the influence of instrument resolution, systematic errors in the experimental data and details of data reduction. More specifically for instrumental resolution, since the comparison is performed on Fourier transformed data rather than acquired data, incorporating the instrument response function is more complex than in the cases mentioned in sections 4.2 and 5. Given that the amount of generated data used to train the models described in this section is large, the reusability of a neural network model trained for a given diffractometer for a different diffractometer is a relevant question. As we are exploring the application of ML to scattering measurements, such questions will become relevant to the community at large.

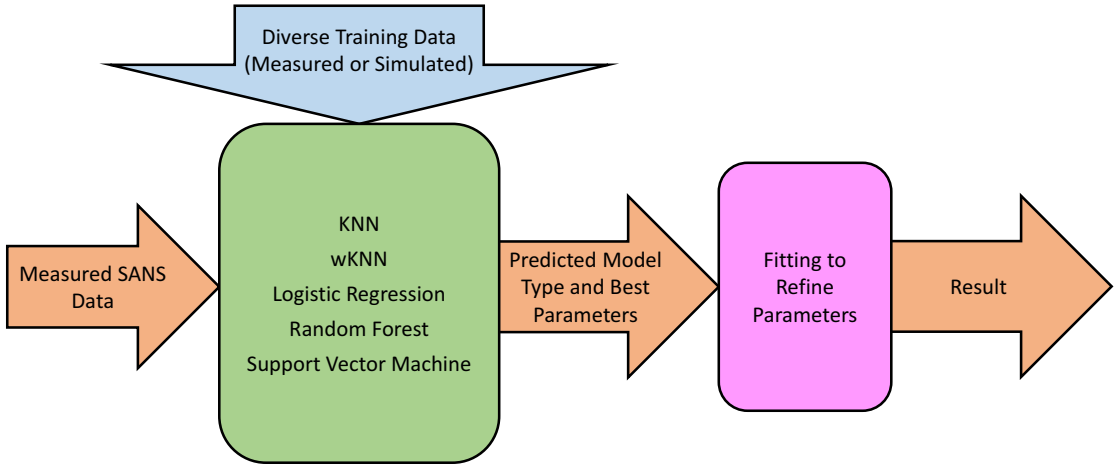

Figure 4. A schematic of the pipeline for using ML methods for classifying SANS data for the best model to use for analysis. The ML can be trained using real or simulated SANS data, as long as the set is diverse. The prediction provided by ML can be fed directly into software for fitting, and it can be used to provide starting parameters for the refinement. The end goal is to provide a useful result to the user [18, 19].

4. Machine learning and SANS

4.1. Classifying SANS data to aid in analysis

One of the greatest challenges that people who use SANS for materials characterization face lies in determining how to fit the data. At the root of the matter is the large number of models that are available for fitting SANS data, such as are implemented in packages such as *SasView* [39] or *SASfit* [40]. The limited information content of SANS data and the potential for models to fit the data well in spite of them often having no relationship with the underlying structure also complicates the matter. As a result, many novice users require assistance from experts to translate measurements into a scientific result that they have confidence in. Two separate efforts have taken place at ORNL to develop ML tools for classifying small-angle scattering data [18, 19].

Archibald and coworkers [18] employed weighted *k*-nearest neighbors (wKNN) [41] for the ML algorithm. wKNN is known as a lazy learning algorithm because it does not create an abstraction of the training data for use when classifying an unknown, such as is done for neural networks. Instead, the unknown is compared against the entire set of training data each time. As a result each classification can be computationally intensive, particularly with large sets of training data. The approach developed employed the euclidean distance, being another form of a least squares comparison, as the criterion for classifying the unknown data set. The performance of the method was quite good, and it was possible to extend the method to perform actual data fitting during classification that was able to inform the user of the model parameters that best fit the data for the selected model. Broadly, the performance of the methods that were trained with and tested against 61 of the models implemented in *SasView* [39] was good against many of the models used, and performance improved when a Gaussian process with or without integrated fitting was employed to expand the capabilities of the classifier [18]. Yet it was clear that many of the closely-related models implemented in *SasView* [39], such as the various models based on spheres or cylinders, present challenges for classification.

In a separate study, Do and coworkers compared several different ML methods for classifying SANS data [19]. In addition to traditional *k*-nearest neighbors (KNN) [42], classifiers based on logistic regression [43], random forest [44], support vector machine [45] and ridge regression [46] approaches were employed. In contrast to wKNN and KNN, these methods translate the data into a model prior to comparing it against a model of the training data set. The authors concluded that the performance of the various methods was comparable.

In both of these studies, the authors used SANS data simulated using *SasView* [39] in the training data set and for testing [18, 19]. In the case of the training data, the use of known models with model parameters that could also be retained with the training data has clear benefits. Foremost, the relationship between the classification and the training data is perfectly clear, which cannot be guaranteed for experimental data. Additionally, a much wider range of model parameters could be simulated than may be available in even the most diverse sets of experimentally measured data. Another potential benefit exists for this approach to training that is less obvious but would be absolutely critical for deploying the method for classifying SANS

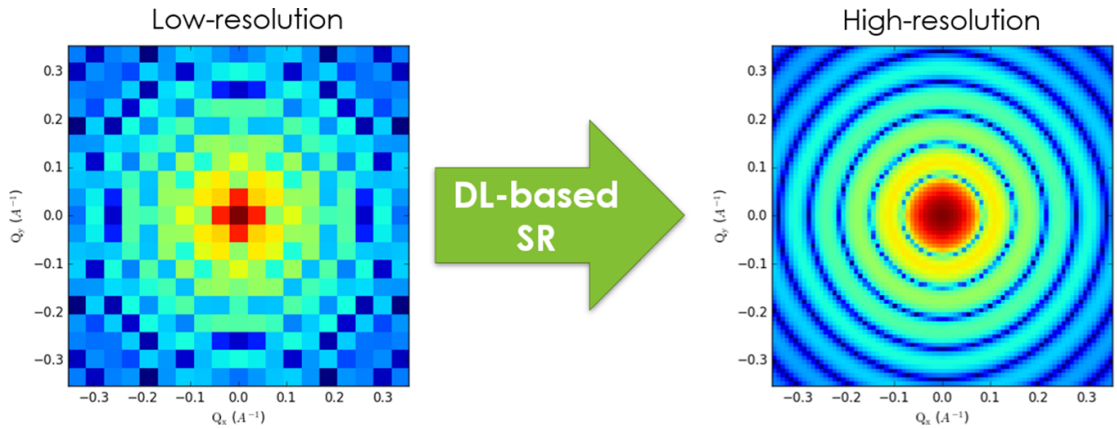

Figure 5. A schematic showing the concept of applying super-resolution algorithm to SANS data. Low-resolution data can be collected at much faster rate by combining detector pixels. The high-resolution data is obtained by using a deep-learning (DL) based super-resolution (SR) algorithm. The scattering data shown here is for a theoretically calculated form factor of a sphere.

data. By training with model intensity profiles that were not convoluted with any instrument resolution information, it is possible to use a single training set to classify data from any configuration of a SANS instrument employed for a measurement. One would simply insert a step for convoluting the training data with the appropriate resolution information prior to the calculation of the distance between the measured and training data.

4.2. Facilitating time-resolved SANS experiments with super resolution techniques

The time-resolved study of materials requires the interpretation of the acquired data at a pace commensurate with the probed time scales. With current instruments, it is possible to design experiments where structural and dynamical changes can be probed *in situ* through controlled changes in sample environment systems [47]. The challenge in this case is to maximize the information one can extract from measurements done on the time scale of seconds with limited neutron flux [48–50]. To achieve a statistically significant result, and in contrast to x-ray measurements, experiments using neutron scattering methods usually require long counting times to compensate for low beam flux and small scattering cross-sections. Novel data analysis tools are therefore necessary to allow real-time characterization of materials using neutron scattering techniques.

For a given collection time, there is a trade-off between counting statistics per detector pixel and detector resolution. One way to speed up the measurement while keeping good counting statistics per detector pixel is to combine detector pixels to increase the effective pixel area, at the expense of detector resolution. Then, obtain the high resolution detector image via super-resolution algorithm. Do and coworkers have adopted a deep-learning based super-resolution algorithm to demonstrate that time-resolved SANS experiments with a finer time resolution can be achieved using neural network models to increase the effective detector resolution [20].

Specifically, scattering data from EQ-SANS [51] at SNS have been selected to train the network model to understand the resolution function of the instrument. Using the ONCat [52] catalog available at ORNL, scattering data from EQ-SANS with identical configuration have been randomly selected. Final training data were further down-selected by choosing data with more than 5 million neutron counts to ensure good statistics of the original resolution. Pairs of high-resolution (original) and low-resolution scattering images were produced by reducing data with different detector binning size. From the low-resolution images, original-resolution scattering images were predicted using both a neural network model and a bicubic interpolation method. The neural network super-resolution method showed much better performance in predicting scattering intensities. Furthermore, the width of scattering peaks was better predicted by the neural network super-resolution model, resulting in more accurate interpretation of low-resolution data. For scattering data without sharp features, both models predicted reasonably well.

Work is in progress at ORNL to improve this super-resolution method. With the existing approach, only data from the same instrument configuration can be processed with confidence. A multi-dimensional parametrization of the resolution deconvolution, similar to one used in super resolution for spectroscopy [53], is being studied to explore the possibility of disentangling the intrinsic detector resolution from the effects of the rest of the instrument configuration. Such an approach would make it possible to assemble training sets from theoretically calculated scattering intensities, which would allow us to use larger

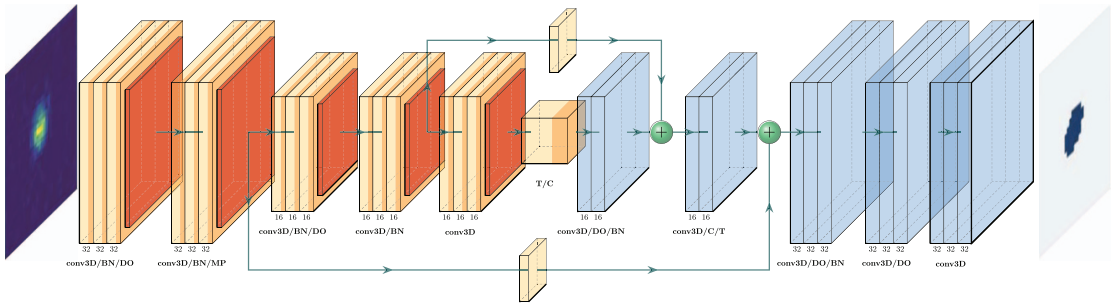

Figure 6. Braggnet processing workflow, showing the mechanisms that take measured data to predict a peak integration mask. The algorithms used to process the data are annotated as follows: three dimensional convolution (conv3D), batch normalization (BN), dropout (DO), max pooling (MP), transpose (T), and concatenate (C).

training sets and potentially develop models targeted to specific materials systems. Developing such tools would greatly expand the range of possible measurements at our SANS instruments.

5. Machine learning to understand diffraction peak shapes for macro molecular crystallography

Crystallography is a fundamental tool in the experimentalist's toolbox that can characterize different materials providing both atomic and molecular structural information [54, 55]. The key to decoding the information from crystallography comes from the analysis of the diffraction pattern produced during experiments. The most dominant features in diffraction patterns are the Bragg peaks' location and intensity. Here we provide a review of Braggnet [21], a ML tool that provides significant improvement to the analysis of Bragg peak location and integrated intensity as compared to existing methods.

Machine learning methods are employed in various related topics in crystallography, such as crystal screening [56, 57], experimental design for protein–drug interactions [58, 59], and crystal detection in single x-ray free-electron laser (XFEL) pulses [60]. However, with the exclusion of Pokric and coworkers' work to learn strong peak shapes using networks of radial basis function, there has been little use of ML to analyze crystallographic data. Braggnet is the first work using ML to addresses all peak shapes and locations, including challenging weak peak identification, and more accurately integrate the scattering intensity in each peak.

The workflow of Braggnet is depicted in figure 6. This figure visualizes the detector locations of a typical crystallography experiment at MaNDi [61]. The upper left image depicts a single Bragg peak being analyzed by Braggnet to predict the peak integration mask used to accurately reconstruct the peak integration value. The improved peak integration and location have a direct impact on the accuracy of the calculation of the nuclear density map and the crystallized protein structure.

The data generation used to train Braggnet is arguably the key to its accuracy and robustness, and we use data augmentation techniques to achieve this goal. The training and testing data is generated based upon fitting profiles of strong experimental peaks to an Ikeda-Carpenter function [62] using *Mantid* [63, 64]. Using these profiles, it is possible to generate a large data set by rotating, translating, and scaling these profiles and adding various strengths of Poisson noise. This method of generating data produces all the different types of Bragg peaks that are expected in experimentation by the underlying physical understanding of crystallography.

The success of Braggnet can be attributed to the network structure and the data construction used to train this ML method. One of the advances of the work of Sullivan and coworkers is that identification of the location and shape of any Bragg peak can be translated into an image segmentation problem. Within the field of image analysis, the problem of image segmentation is to identify all pixels in an image associated with a particular object. For example identifying all the pixels that depict cows in a picture of a farm. There is a rich and ongoing field of research for image segmentation, and ML has recently dominated this field with accurate segmentation of images using techniques such as encoder–decoder architectures (U-Nets) [65], dilation [66], and advanced pooling (down-sampling) schemes [67]. Braggnet uses a convolutional U-Nets architecture that has been demonstrated to segment volumetric images accurately using a sparse set of training data, with the added advantage of demonstrated generalizability.

Conclusion

The wide variety of science that can be tackled with neutron scattering continually creates opportunities for developing new data analysis methods. While expanding the repertoire of modeling tools, ML techniques offer opportunities to address widespread issues with the analysis of neutron data. The challenges of analyzing scattering data are such that researchers wanting to model their data often require the expertise of scattering scientists. But even for practitioners, extracting structural information from neutron data can be a time consuming endeavor. In some cases, traditional regression approaches are simply insufficient and new techniques need to be explored.

Here, we described applications of ML for working with neutron scattering data. Samarakoon and coworkers [12] demonstrated how they could extract solutions to magnetic diffuse scattering that would otherwise be difficult and time consuming. Garcia-Cardona [17] and coworkers applied ML approaches to predicting structures from powder data. Archibald and coworkers [18] applied KNN algorithms to assist users with the task of determining the most appropriate model for fitting SANS data. Do and coworkers [19] used KNN and other methods to do the same. ML has also been shown to help address instrument effects. Chang and coworkers [20] applied super-resolution techniques to enhance SANS data in view of enabling finer time-resolved experiments. Sullivan and coworkers [21] have applied ML to automatically extract Bragg peak parameters in diffraction data. These efforts are part of a growing community emphasis on using ML techniques to support scientific user facilities. Together, this work demonstrates the potential of ML for speeding up several processes in the experiment life cycle at user facilities. Although these address immediate needs, they will also pave the way to a more integrated approach to science where data analytics workflows can be used to combine scientific data from different measurement techniques to accelerate scientific productivity.

Acknowledgments

A portion of this research used resources at the SNS, a Department of Energy (DOE) Office of Science User Facility operated by ORNL. Part of the research was sponsored by ExaLearn, an Exascale Computing Project, DOE. ORNL is managed by UT-Battelle LLC for DOE under Contract DE-AC05-00OR22725. We acknowledge the support by the Scientific Discovery through Advanced Computing (SciDAC) funded by U S Department of Energy, Office of Science, Advanced Scientific Computing Research through FASTMath Institutes. Portions of the diffuse scattering and SANS research were sponsored by the Laboratory Directed Research and Development Program of ORNL, managed by UT-Battelle, LLC, for the U S Department of Energy. The DTO work was run on the HPC resources of the OLCF. We acknowledge the contributions of B Sullivan, K Barros, C Batista, V Lynch, P Langan. We had useful discussion with J Y Y Lin, and Y-M Lee.

Data availability statement

No new data were created or analyzed in this study.

ORCID iDs

Mathieu Doucet https://orcid.org/0000-0002-5560-6478
D Alan Tennant https://orcid.org/0000-0002-9575-3368
Garrett E Granroth https://orcid.org/0000-0002-7583-8778

References

[1] Perrault R, Shoham Y, Brynjolfsson E, Clark J, Etchemendy J, Grosz B, Lyons T, Manyika J, Mishra S and Niebles J C 2019 *The AI Index 2019 Annual Report*
[2] Abadi M *et al* 2016 TensorFlow: large-scale machine learning on heterogeneous distributed systems Software available from (https://www.tensorflow.org/)
[3] Pedregosa F *et al* 2011 Scikit-learn: machine learning in python *J. Mach. Learn. Res.* **12** 2825–30
[4] Chollet Fçois *et al* Keras (https://keras.io)
[5] Ratner D, Sumpter B and Alexander F *et al* 2019 BES roundtable on producing and managing large scientific data with artificial intelligence and machine learning *OSTI Technical Report* (https://doi.org/10.2172/1630823)
[6] Herwig K W (ed) 2020 First experiments: new science opportunities at the spallation neutron source second target station *Technical Report* ORNL/SPR-2020/1437
[7] Fagnan K, Nashed Y, Perdue G, Ratner D, Shankar A and Yoo S 2019 Data and models: a framework for advancing AI in science *OSTI Technical Report* (https://doi.org/10.2172/1579323)
[8] Mason T E *et al* 2006 The spallation neutron source in oak ridge: a powerful tool for materials research *Physica* B **385** 955–60

[9] Squires G L 1978 *Introduction to the Theory of Thermal Neutron Scattering* (Cambridge: Cambridge University Press)
[10] Sands D E 1993 *Introduction to Crystallography* Dover Books on Chemistry (New York: Dover Publications) p 103
[11] Sivia D S 2011 *Elementary Scattering Theory: For X-Ray and Neutron Users* (Oxford: Oxford University Press) p 48
[12] Samarakoon A M *et al* 2020 Machine-learning-assisted insight into spin ice $Dy_2Ti_2O_7$ *Nat. Commun.* **11** 892
[13] Littlewood P and Proffen T 2015 Frontiers in data, modeling, and simulation (grand challenges workshop report) (https://neutrons.ornl.gov/sites/default/files/FrontiersInData-WorkshopReport-Mar2015.pdf)
[14] 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** 1342–7
[15] Carmona-Loaiza J M 2020 Towards reflectivity profile inversion through artificial neural networks (https://arxiv.org/abs/2010.07634)
[16] Doucet M, Archibald R K and Heller W T 2020 (submitted for publication)
[17] Garcia-Cardona C, Kannan R, Johnston T, Proffen T, Page K and Seal S K 2019 Learning to predict material structure from neutron scattering data *2019 IEEE Int. Conf. Big Data (Big Data)* pp 4490–7
[18] Archibald R K, Doucet M, Johnston T, Young S R, Yang E and Heller W T 2020 Classifying and analyzing small-angle scattering data using weighted *k* nearest neighbors machine learning techniques *J. Appl. Crystallogr.* **53** 326–34
[19] Changwoo D, Chen W-R and Lee S 2020 Small angle scattering data analysis assisted by machine learning methods *MRS Adv.* **5** 1577–84
[20] Chang M-C, Wei Y, Chen W-R and Do C 2019 Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning (http://arxiv.org/abs/1904.08450)
[21] Sullivan B, Archibald R, Azadmanesh J, Vandavasi V G, Langan P S, Coates L, Lynch V and Langan P 2019 BraggNet: integrating Bragg peaks using neural networks *J. Appl. Crystallogr.* **52** 854–63
[22] McGreevy R L 2001 Reverse Monte Carlo modelling *J. Phys.: Condens. Matter* **13** R877–913
[23] Lovesey S W 1984 *Theory of Neutron Scattering from Condensed Matter* vol 2 (Oxford: Oxford University Press)
[24] Stone M B *et al* 2014 A comparison of four direct geometry time-of-flight spectrometers at the spallation neutron source *Rev. Sci. Instrum.* **85** 045113
[25] Toth S and Lake B 2015 Linear spin wave theory for single-q incommensurate magnetic structures *J. Phys.: Condens. Matter* **27** 166002
[26] Hahn S E *et al* 2018 Spinwavegenie (https://github.com/SpinWaveGenie/SpinWaveGenie)
[27] Hey T, Butler K, Jackson S and Thiyagalingam J 2020 Machine learning and big scientific data *Philos. Trans. R. Soc.* A **378** 20190054
[28] Huberman T, Tennant D A, Cowley R A, Coldea R and Frost C D 2008 A study of the quantum classical crossover in the spin dynamics of the 2d $s = 5/2$ antiferromagnet Rb_2MnF_4: neutron scattering, computer simulations and analytic theories *J. Stat. Mech.: Theory Exp.* **2008** P05017
[29] Samarakoon A M, Banerjee A, Zhang S-S, Kamiya Y, Nagler S E, Tennant D A, Lee S-H and Batista C D 2017 Comprehensive study of the dynamics of a classical Kitaev spin liquid *Phys. Rev.* B **96** 134408
[30] Samarakoon A M, Wachtel G, Yamaji Y, Tennant D A, Batista C D and Kim Y B 2018 Classical and quantum spin dynamics of the honeycomb Γ model *Phys. Rev.* B **98** 045121
[31] Oak Ridge National Laboratory 2020 Oak Ridge Leadership Computing Facility (https://www.olcf.ornl.gov/)
[32] Feng Y, Liu Y, Whitfield R, Osborn R and Rosenkranz S 2018 Implementation of cross correlation for energy discrimination on the time-of-flight spectrometer CORELLI *J. Appl. Crystallogr.* **51** 315–22
[33] Zhang W, Wang X, Zhao D and Tang X 2012 Graph degree linkage: agglomerative clustering on a directed graph *Computer Vision—ECCV 2012* eds A Fitzgibbon, S Lazebnik, P Perona, Y Sato and C Schmid (Berlin, Heidelberg: Springer) pp 428–41
[34] Liu C-H, Tao Y, Hsu D, Du Q and Billinge S J L 2019 Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function *Acta Crystallogr.* A **75** 633–43
[35] Th P, S Egami B, T and Louca D 2003 Structural analysis of complex materials using the atomic pair distribution function—a practical guide *Z. Krist.* **218** 132–43
[36] Rietveld H M 1969 A profile refinement method for nuclear and magnetic structures *J. Appl. Crystallogr.* **2** 65–71
[37] Toby B H and Von Dreele R B 2013 GSAS-II: the genesis of a modern open-source all purpose crystallography software package *J. Appl. Crystallogr.* **46** 544–9
[38] Neuefeind J, Feygenson M, Carruth J, Hoffmann R and Chipley K K 2012 The nanoscale ordered materials diffractometer nomad at the spallation neutron source SNS *Nucl. Instrum. Methods Phys. Res.* B **287** 68–75
[39] Doucet M *et al* 2018 Sasview version 4.2 10.5281/zenodo.1412041 sasview.org
[40] Breßler I, Kohlbrecher J and Thünemann A F 2015 *SASfit*: a tool for small-angle scattering data analysis using a library of analytical expressions *J. Appl. Crystallogr.* **48** 1587–98
[41] Wettschereck D, Aha D W and Mohri T 1997 A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms *Artif. Intell. Rev.* **11** 273–314
[42] Altman N S 1992 An introduction to kernel and nearest-neighbor nonparametric regression *Am. Stat.* **46** 175–85
[43] Berkson J 1950 Are there two regressions? *J. Am. Stat. Assoc.* **45** 164–80
[44] Ho T K 1998 The random subspace method for constructing decision forests *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 832–44
[45] Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97
[46] Saunders C, Gammerman A and Vovk V 1998 Ridge regression learning algorithm in dual variables *Proc. Fifteenth Int. Conf. Machine Learning ICML '98* (San Francisco, CA: Morgan Kaufmann Publishers Inc) p 515–21
[47] Granroth G E *et al* 2018 Event-based processing of neutron scattering data at the spallation neutron source *J. Appl. Crystallogr.* **51** 616–29
[48] Lund R, Willner L, Richter D, Iatrou H, Hadjichristidis N, Lindner P and IUCr 2007 Unraveling the equilibrium chain exchange kinetics of polymeric micelles using small-angle neutron scattering—architectural and topological effects *J. Appl. Crystallogr.* **40** s327–31
[49] Bruetzel L K, Walker P U, Gerling T, Dietz H and Lipfert J 2018 Time-resolved small-angle x-ray scattering reveals millisecond transitions of a DNA origami switch *Nano Lett.* **18** 2672–6
[50] Sauter A, Roosen-Runge F, Zhang F, Gudrun Lotze R M Jacobs J and Schreiber F 2015 Real-time observation of nonclassical protein crystallization kinetics *J. Am. Chem. Soc.* **137** 1485–91
[51] Heller W T *et al* 2018 The suite of small-angle neutron scattering instruments at Oak Ridge National Laboratory *J. Appl. Crystallogr.* **51** 242–8
[52] Parker P G and Ren S 2019 ONCat (ORNL Neutron Catalog) (https://doi.org/10.11578/dc.20200513.5)

[53] Islam F, Lin J Y Y, Archibald R, Abernathy D L, Al-Qasir I, Campbell A A, Stone M B and Granroth G E 2019 Super-resolution energy spectra from neutron direct-geometry spectrometers *Rev. Sci. Instrum.* **90** 105109
[54] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H N S I, Shindyalov I N and Bourne P E 2000 The protein data bank *Nucl. Acids Res.* **28** 235–42
[55] Groom C R, Bruno I J, Lightfoot M P and Ward S C 2016 The Cambridge structural database *Acta Crystallogr.* B **72** 171–9
[56] Liu R, Freund Y and Spraggon G 2008 Image-based crystal detection: a machine-learning approach *Acta Crystallogr.* D **64** 1187–95
[57] Bruno A E, Charbonneau P, Newman J, Snell E H, So D R, Vanhoucke V, Watkins C J, Williams S and Wilson J 2018 Classification of crystallization outcomes using deep convolutional neural networks *PLoS ONE* **13** 1–16
[58] Zhang L, Tan J, Han D and Zhu H 2017 From machine learning to deep learning: progress in machine intelligence for rational drug discovery *Drug Discov. today* **22** 1680–5
[59] Ding H, Takigawa I, Mamitsuka H and Zhu S 2013 Similarity-based machine learning methods for predicting drug–target interactions: a brief review *Brief. Bioinform.* **15** 734–47
[60] Ke T W, Brewster A S, Yu S X, Ushizima D, Yang C and Sauter N K 2018 A convolutional neural network-based screening tool for x-ray serial crystallography *J. Synchrotron Radiat.* **25** 655–70
[61] Spallation Neutron Source 2020 The macromolecular neutron diffractometer (MANDI) (https://neutrons.ornl.gov/mandi)
[62] Ikeda S and Carpenter J M 1985 Wide-energy-range, high-resolution measurements of neutron pulse shapes of polyethylene moderators *Nucl. Instrum. Methods Phys. Res.* A **239** 536–44
[63] Arnold O *et al* 2014 Mantid—data analysis and visualization package for neutron scattering and μ sr experiments *Nucl. Instrum. Methods Phys. Res.* A **764** 156–66
[64] Sullivan B *et al* Improving the accuracy and resolution of neutron crystallographic data by three-dimensional profile fitting of Bragg peaks in reciprocal space *Acta Crystallogr.* D 74 1085–95 2018
[65] Ronneberger O, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* eds N Navab, J Hornegger, W M Wells and A F Frangi (Cham: Springer Int. Publishing) pp 234–41
[66] Fisher Y and Koltun V 2016 Multi-scale context aggregation by dilated convolutions (arXiv: 1511.07122)
[67] Zhao H, Shi J, Qi X, Wang X and Jia J 2017 Pyramid scene parsing network *2017 Conf. Computer Vision and Pattern Recognition (CVPR)* pp 6230–9