Taylor & Francis
Taylor & Francis Group

Synchrotron Radiation News

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gsrn20

Challenges and Opportunities of Machine Learning on Neutron and X-ray Scattering

Nathan C. Drucker, Tongtong Liu, Zhantao Chen, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Thanh Nguyen, Yao Wang & Mingda Li

To cite this article: Nathan C. Drucker, Tongtong Liu, Zhantao Chen, Ryotaro Okabe, Abhijatmedhi Chotrattanapituk, Thanh Nguyen, Yao Wang & Mingda Li (2022) Challenges and Opportunities of Machine Learning on Neutron and X-ray Scattering, Synchrotron Radiation News, 35:4, 16-20, DOI: 10.1080/08940886.2022.2112498

To link to this article: https://doi.org/10.1080/08940886.2022.2112498

Published online: 12 Oct 2022.

OPEN ACCESS

Feature Article

Challenges and Opportunities of Machine Learning on Neutron and X-ray Scattering

Nathan C. Drucker,[1,2] Tongtong Liu,[1,3] Zhantao Chen,[1,4] Ryotaro Okabe,[1,5]
Abhijatmedhi Chotrattanapituk[1,6] Thanh Nguyen,[1,7] Yao Wang,[8] and Mingda Li[1,7]

[1]*Quantum Measurement Group, MIT, Cambridge, Massachusetts, USA*
[2]*School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA*
[3]*Department of Physics, MIT, Cambridge, Massachusetts, USA*
[4]*Department of Mechanical Engineering, MIT, Cambridge, Massachusetts, USA*
[5]*Department of Chemistry, MIT, Cambridge, Massachusetts, USA*
[6]*Department of Electrical Engineering and Computer Science, MIT, Cambridge, Massachusetts, USA*
[7]*Department of Nuclear Science and Engineering, MIT, Cambridge, MA, USA*
[8]*Department of Physics and Astronomy, Clemson University, Clemson, South Carolina, USA*

Introduction

Machine learning has been highly successful in boosting the research for neutron and X-ray scattering in the past few years [1, 2]. For diffraction, machine learning has shown great promise in phase mapping [3, 4] and crystallographic information determination [5, 6]. In small-angle scattering, machine learning shows the power in reaching super-resolution [7, 8], reconstructing structures for macromolecules [9], and building structure-property relations [10]. As for absorption spectroscopy, machine learning has enabled the rapid inverse search for optimized structures [11, 12] with improved spectral interpretability [13, 14]. Overall, as a data-driven approach, the success of the machine-learning-based scattering analysis depends on a few criteria, including:

- Quantity of available experimental data, and feasibility to extract certain data labels;
- Quality of experimental data that can separate the intrinsic effect (e.g., materials properties) from extrinsic influence (e.g., instrumental or data artifacts);
- Feasibility to generate high volume of computational data;
- Accuracy of computational data that can simulate the experimental data.

Based on these criteria, it is understandable that not all scattering techniques are equally feasible for carrying out machine learning. For instance, small-angle scattering and X-ray absorption (XAS) have become two frontiers in applying machine learning techniques, thanks to the low data dimension (1D data), relatively low technical barrier on performing measurements, and high feasibility to generate a large volume of computational data that can faithfully represent the experiments.

In general, low-dimensional data are relatively easier to use for machine learning training. However, some of the most powerful scattering techniques lie in a higher data dimension, such as inelastic scattering, which resides in the 4D momentum–energy (k × E) space. This poses a dilemma in using machine learning on scattering spectroscopies: the more powerful a scattering technique is, the more likely the technique is in a higher data dimension, and the more challenge it will face in performing machine learning. Table 1 lists a few common scattering and spectroscopy techniques that are ranked by the feasibility to perform machine learning, based on the above criteria.

From Table 1 we see that the *t*-axis generally increases the challenge of machine learning training with higher dimensional data to analyze but lower amount of available training data. To address this challenge, it is worthwhile mentioning that even without any training data, a branch of machine learning, termed scientific machine learning [15], can be adopted to perform time-resolved data analysis. In one work by Chen et al. [16], it is shown that scientific machine learning has broken the challenges to investigate frequency-resolved phonon thermal transport by analyzing time-resolved diffraction patterns. By making an assumption that the phonon transport can be described by the Boltzmann transport equation (BTE), and linking the atomic displacements to the Debye-Waller smearing of diffraction intensities, it shows the possibility to acquire high-dimensional frequency-dependent thermal transport from time evolution of diffraction intensities, mainly phonon relaxation times and interfacial thermal transmission coefficients. This enables a direct reconstruction of real-space, real-time, frequency-resolved phonon dynamics across an interface of the heterostructure with sub-ps resolution. Given the fact that many time-resolved scattering techniques are described by a certain dynamical equation, scientific machine learning is anticipated to play an increasingly important role in analyzing time-resolved scattering data.

Beyond that, even at low data dimensions, there are still outstanding problems that exist. In this perspective, we introduce three such problems, including the structure determination with defects or magnetism, the learning of magnetic excitations, such as phonons and magnons, and the prediction of the microscopic interaction in strongly correlated electron systems.

Table 1: List of common scattering and spectroscopic techniques and their ranked feasibility to perform machine learning, with easy (green), medium (yellow), and hard (brown) levels.

Space	Neutron	X-ray
r	🟨 Neutron imaging	🟨 X-ray microscopy
k	🟩 (Polarized) reflectometry (PNR)	🟩 X-ray reflectometry (XRR)
	🟩 Small-angle neutron scattering (SANS)	🟩 Small-angle X-ray scattering (SAXS)
	🟨 Neutron diffraction	🟨 X-ray diffraction (XRD)
E	🟨 Vibrational spectroscopy	🟩 X-ray absorption (XAS)
		🟫 Resonant inelastic X-ray scattering (RIXS)
$k \times t$	🟫 Time-resolved diffraction/scattering	🟫 X-ray photon correlation spectroscopy (XPCS)
$E \times k$	🟫 Inelastic neutron scattering (INS)	🟫 Angular-resolved photoemission (ARPES)
	🟫 Quasi-elastic neutron scattering (QENS)	🟫 Non-resonant inelastic X-ray scattering (IXS)
$E \times t$	N/A	🟨 Coherent diffraction imaging
$E \times k \times t$	N/A	🟫 Time-resolved (tr) XAS, tr-RIXS

Defect and magnetic structure determination

The structure determination for small-angle scattering (including SANS and SAXS) and reflectometry (including NR and XRR) is relatively straightforward for three reasons. First, the data dimension is low. Except for some rare cases, such as the magnetic vortices lattice using SANS, which has 2D data [17, 18], most of the small-angle scattering and reflectometry data can be considered as 1D curves after data reduction. Second, the computational cost to generate the spectra is low, without the need of atomistic-scale structure details. Third, the computational data can faithfully represent the measured experimental data directly. This is because small-angle scattering and reflectometry probe the course-grained, nanoscale structures, where the structure parameters in the materials are abstracted into a few parameters.

However, significant challenges still remain in X-ray and neutron diffraction. As a probe that detects the atomistic-scale structure, diffraction data is generally 3D in nature. Moreover, the atomic structure has a design space that can grow exponentially with the unit-cell size. In fact, the number of possible structure C for a unit cell with N atoms and volume V can be written as [19]

$$C=\frac{1}{(V/\delta^3)}\frac{\left(V/\delta^3\right)!}{\left[\left(V/\delta^3\right)-N\right]!N!} \qquad (1)$$

where $\delta \sim 1\text{Å}$ is the discretization parameter. This can easily lead to astronomically large combinations of possible structures. Even for potentially stable structures with structure relaxation, the number of local minimal energy configurations still grows exponentially with N.

Besides the structure combinatorics, even within a fixed crystalline solid, the structure determination can be significantly hampered due to the crystallographic defects. The point defects will slightly change the unit cell but only lead to small weight in diffraction patterns [20], particularly at the low-concentration regime, while the line defects and planar defects such as dislocations and grain boundaries can be considered a multiscale problem that will alter both atomistic and mesoscopic structures. Analyzing the defect information beyond current refinement schemes poses another grand challenge.

To address the grand challenge, efficient generation of materials provides an alternative but promising approach through generative models. Rather than conventional supervised learning that validates targeted materials property by inputting a fixed material, in generative models, new molecules and new crystals can be generated directly from, say, random noises after the learning stage. Generative adversarial networks and variational autoencoders represent two approaches of generative models, where the former performs discriminator operation against the “real data”, i.e., the training set, while the latter perform one-shot generation through the latent space encoding and latent space manipulation. One of the more recent developments is the development of diffusion models [21]. In a diffusion model, the data are gradually turned into noise, and neural networks are used to invert the procedure in a step-by-step manner, where each step of inversion, the data becomes less noisy. The diffusion models allow the fine tuning at each step and gains flexibility toward constrained optimization. Recently, the diffusion model has been implemented in conjunction with variational autoencoder to generate crystalline materials with targeted performance [22].

Beyond atomic and defect structure determination, a third challenge here is magnetic structure determination. Even with a well-defined atomic structure, adding magnetic degrees of freedom at least adds the spin vectors to the magnetic ions onto a given magnetic atom. The existence of the rich magnetic structures with larger magnetic unit cells, nonzero magnetic propagation vectors, and incommensurate magnetic structure all add significant complexity to the solution space of magnetic structures.

Overall, despite the straightforward calculation from atomic or magnetic configurations to diffraction patterns, the inverse problem of solving atomic, defect, and magnetic structure has posed a grand challenge due to the large dimension of the parameter space.

To tackle the challenges, for the defect structures, it is possible to only focus on one of a few materials and span the defect parameter space instead of starting from a universal defect predictor. Besides, since the structure factor S(Q) for a perfect crystal is a series of delta-function, while the structure factor for disordered solid is continuous in the reciprocal space, it is crucial to find out the proper defect descriptor in reciprocal space or other latent spaces. For instance, the environment in grain boundaries can be captured through the features from the smooth overlap of atomic positions [23]. As to the magnetic structures, it is important to utilize the experimental magnetic structure data as the training set and adopt approaches that can augment the data set, such as through crystallographic symmetry [24].

Learning and predicting elementary excitations

One of the main powers of neutron and X-ray scattering measurements is to directly measure elementary excitations, such as phonons [25, 26], magnons [27, 28], and spinons [29, 30]. Compared to diffraction problems where the solution space of the inverse problem is huge, for elementary excitations, the main bottleneck lies in the huge computational cost even for the forward problem. For instance, to perform the phonon dispersion calculation with first-principles density functional perturbation theory (DFPT), the computational cost C is on the order of [31]

$$C \sim R_{IFC}^{3} \times 3N^{4} \tag{2}$$

where $R_{IFC} \sim 2\pi / \Delta q$ is the interatomic-spacing range, factor “3” accounts for the three phonon polarization modes, and N is the number of atoms in the unit cell.

Machine learning methods can be used to accurately predict phonon bandstructures without the high computational cost of *ab initio* methods. One method is the Gaussian approximation potential (GAP) [32], which has shown to be accurate in predicting the lattice excitations of crystalline phases in addition to phases that may have vacancies or other defects [33]. This machine learning-based method is faster than typical *ab initio* methods, and more accurate than empirical potential (EP) methods [34]. Beyond the GAP approach, by using the crystallographic symmetry and symmetry-preserved neural networks, Chen and Andrejevic et al. demonstrated a direct prediction of phonon density-of-states from atomic coordinates [35]. With faster simulation of phonons, it will become more feasible to train machine learning models to identify and predict lattice excitations.

There is also a large computational cost to predicting magnetic excitations in quantum magnets. Even a relatively simple Hamiltonian for a magnetic system with exchange interaction such as

$$H = \sum\nolimits_{n} J_{n} \sum\nolimits_{<i,j>_{n}} S_{i} \bullet S_{j} \tag{3}$$

where the sum includes up to nth-nearest-neighbor exchange coupling parameters J_n can lead to intricate magnetic excitations. Many magnetic materials also include additional terms, such as magnetic anisotropy, dipolar interactions, and the Dzyaloshinskii-Moriya interaction. Each of these interactions serves as an additional dimension in the parameter space that needs to be accounted for in simulations in the forward problem, and machine learning in the inverse problem. Once a model Hamiltonian is identified, a common method of simulating measurable quantities is with Monte Carlo (MC) methods. For each (classical) MC simulation, a particular set of randomized parameters is selected, but then the simulation must relax from an initial configuration to the final ground state that will eventually be compared to experiments. For example, Samarakoon et al. use 1000 Monte Carlo simulations of a magnetic Hamiltonian, which includes three nearest-neighbor couplings and dipolar interactions to gain insight into spin ice $Dy_2Ti_2O_7$ by using an autoencoder [36].

One way to approach the challenge of generating a large enough number of magnetic system simulations to feed into a machine learning model is to actually use machine learning to accelerate the Monte Carlo simulations themselves [37, 38]. Liu and Qi et al. approach speed up their magnetic simulation by an order of magnitude through the use of a “self-learning” Monte Carlo method [37]. This method runs a simulation of a local update and then learns trends about the update process to guide global configuration updates. By accelerating simulations of spin Hamiltonians, it has shown great power in predicting regions near the phase transition, and will become easier to extract parameters from experimental data and predict materials’ magnetic excitations.

Prediction of strongly correlated systems

Another challenge that is crucial to tackle but rarely accomplished is to predict strongly correlated systems with machine learning. Strongly correlated systems [39] are highly nontrivial due to the interplay between charge, spin, orbital, and lattice degrees of freedom; the computational power required for predicting a strongly correlated system usually grows exponentially with system size (see Table 2 for more details). Experimental measurements are restricted to limited observables in the vast Hilbert space and the inverse problem is challenging for it is a task of solving a system of exponentially growing underlying dimensions with observables that grow only polynomially with system size.

Table 2. Common computational approaches for strongly correlated systems.

		Essence	Pros	Cons
DFT-based method	DFT + U	Pseudopotential + correction	• Straightforward correction to first-principles simulations • Low computational cost O(N^3)	- Lack of static correlation - Poor experimental comparison
Green's function perturbation theory	GW	Self-consistent correction of Green's function	• Capturing correlations beyond the static limit • Portable to DFT and wavefunction methods	- Complication for multi-particle excitations - Restricted to weakly correlated or high-dimensional systems
	DMFT			
Wavefunction-based method	Hatree-Fock and post-HF methods	Exact or variational wavefunctions and density matrix in ensembles	• Capturing all correlation and entanglement effects allowed in a given subspace • Good mathematical structure for variational principles • Some methods are accurate or asymptotically accurate	- Exponential complexity (at least superlinear) - Restricted to small or low-dimensional systems
	Exact Diagonalization			
	DMRG and Tensor network			
Monte Carlo	Quantum Monte Carlo	Importance sampling of configurations or auxiliary fields	• Accurate except for a statistical error • Polynomial scaling with system size	- Inefficient at low temperatures - Restricted to thermal equilibrium - Fermion-sign problem for (quantum) fermionic systems
	Classical Monte Carlo			

On the machine learning side, even analyzing polynomially growing experimental data is challenging since they are high-dimensional in nature. Many powerful methods to measure strongly correlated systems, such as RIXS, INS, APRES, even tr-ARPES, tr-RIXS, are very high-dimensional; for example, INS of three-dimensional materials are four dimensions in momentum–energy ($k \times E$) space, while tr-ARPES in momentum–energy–time ($k \times E \times t$) space. This posed an intrinsic challenge for both data processing and inverse problem. On the computation side, there are some external challenges in spectra calculation of non-equilibrium systems, while pump-probe techniques that drive the system out of equilibrium and induce collective excitations play important roles in the experimental studies of strongly correlated systems. Taking tr-RIXS for example, in order to mimic the resonant scattering process and explore rich physics such as multi-particle excitations, one needs to take into account the finite lifetime of the intermediate state and higher-order correlations beyond linear response [40], leading to a computational complexity of $\mathrm{O}\left(N_t^4\right)$, where N_t is the number of evolution steps in time. Nevertheless, there are increasing developments [41] in the theory of calculating tr-RIXS [42] and tr-ARPES [43] spectra. These simulation results can benefit the training process and be extended to analyze real experimental data.

To tackle the challenge, instead of spanning the large parameter space over all possible regions, what can be done is to focus on one material, and span the parameter space through external control knobs. This follows the same philosophy as current research on machine learning of strongly correlated systems, which usually focuses on a simplified toy model with only a few parameters, while targeting the spectra of one realistic material can lead to more practical applications. More importantly, one can perform joint analysis with different kinds of spectra; for example, learning the effective interaction from both INS and APPES data. This effectively enlarges the information contained in the input data with a relatively small cost, since different spectra contain information on different degrees of freedom. Such study is still in its infancy and is highly promising in the future.

Outlook

In this perspective, we introduced three categories of materials research problems, which may meet significant challenges for machine learning, but may also benefit most when performing machine learning properly. The three problems, predicting structures, predicting elementary excitations, and predicting correlated systems, are all inverse problems that aim to learn materials information from neutron and X-ray scattering data. However, they differ from the forward problems:

Structure prediction: efficient and reliable forward calculations, (but exponentially large space).
Elementary excitation prediction: inefficient but reliable forward calculations.
Correlated system prediction: inefficient and less-reliable forward calculations.

In light of this, running more forward calculations, adopting a more efficient way to run forward calculations, and new ways to perform measurements to learn more information, become the natural and potentially unavoidable setups to drive the fields forward.

Acknowledgments

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

Z.C., T. L., and M.L. acknowledge support from United States Department of Energy (DOE), Basic Energy Sciences Award No. DE-SC0020148 and DE-SC0021940. N.A. acknowledges National Science Foundation (NSF) Graduate Research Fellowships Program support under Grant No. 1122374. M. L. acknowledges support from Norman C. Rasmussen Career Development Chair. ■

References

1. Z. Chen et al., *Chem. Phys. Rev.* 2 (3), 031301 (2021). doi:10.1063/5.0049111
2. M. Doucet et al., *Mach. Learn: Sci. Technol.* **2** (2), 023001 (2021). doi:10.1088/2632-2153/abcf88
3. J. Bai et al., *AI Mag.* **39** (1), 15 (2018). doi:10.1609/aimag.v39i1.2785
4. V. Stanev et al., *NPJ Comput. Mater.* **4** (1), (2018). doi:10.1038/s41524-018-0099-2
5. F. Oviedo et al., *NPJ Comput. Mater.* **5** (1), (2019). doi:10.1038/s41524-019-0196-x
6. C. H. Liu et al., *Acta Crystallogr. A Found. Adv.* **75** (4), 633 (2019). doi:10.1107/S2053273319005606
7. M.-C. Chang et al., *MRS Commun.* **10** (1), 11 (2020). doi:10.1557/mrc.2019.166
8. C. Do et al., *MRS Adv.* **5** (29–30), 1577 (2020). doi:10.1557/adv.2020.130
9. H. He et al., *iScience* **23** (3), 100906 (2020). doi:10.1016/j.isci.2020.100906
10. E. Y. Lee et al., *Proc. Natl. Acad. Sci. USA.* **113** (48), 13588 (2016). doi:10.1073/pnas.1609893113
11. J. Timoshenko and A. I. Frenkel, *ACS Catal.* **9** (11), 10192 (2019). doi:10.1021/acscatal.9b03599
12. A. A. Guda et al., *Catal. Today* **336**, 3 (2019). doi:10.1016/j.cattod.2018.10.071
13. M. R. Carbone et al., *Phys. Rev. Lett.* **124** (15), 156401 (2020). doi:10.1103/PhysRevLett.124.156401
14. M. R. Carbone et al., *Phys. Rev. Mater.* **3** (3), 033604 (2019). doi:10.1103/PhysRevMaterials.3.033604
15. C. Rackauckas et al., Universal differential equations for scientific machine learning (2021). doi:10.48550/arXiv.2001.04385
16. Z. Chen et al., Panoramic mapping of phonon transport from ultrafast electron diffraction and machine learning (2022). doi:10.48550/arXiv.2202.06199
17. S. Mühlbauer et al., *Rev. Mod. Phys.* **91**, 015004 (2019). doi:10.1103/RevModPhys.91.015004
18. Y. Li et al., *Phys. Rev. B.* **83** (5), (2011). doi:10.1103/PhysRevB.83.054507
19. A. R. Oganov et al., *Nat. Rev. Mater.* **4** (5), 331 (2019). doi:10.1038/s41578-019-0101-8
20. J. Als-Nielsen and D. McMorrow, *Elements of Modern X-Ray Physics*, 2nd ed. (Wiley, New York, 2011).
21. M. Wagih et al., *Nat. Commun.* **11** (1), 6376 (2020). doi:10.1038/s41467-020-20083-6
22. T. E. Smidt et al., *Phys. Rev. Research* **3**, L012002 (2021).
23. Y. Song and S. Ermon, Generative modeling by estimating gradients of the data distribution (2020). doi:10.48550/arXiv.1907.05600
24. T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, Crystal diffusion variational autoencoder for periodic material generation (2022). doi:10.48550/arXiv.2110.06197
25. J. L. Niedziela et al., *Nat Phys*, **15**, 73 (2018). doi:10.1038/s41567-018-0298-2
26. T. Nguyen et al., *Phys. Rev. Lett.* **124** (23), 236401 (2020). doi:10.1103/PhysRevLett.124.236401
27. Y. Li et al., *Nature* **468** (7321), 283 (2010). doi:10.1038/nature09477
28. I. A. Zaliznyak and J. M. Tranquada, *Strongly Correlated Systems: Experimental Techniques*, edited by A. Avella and F. Mancini (Springer, Berlin, 2015), 205–235.
29. A. Banerjee et al., *Science* **356** (6342), 1055 (2017). doi:10.1126/science.aah6015
30. Y. Shen et al., *Nature* **540** (7634), 559 (2016). doi:10.1038/nature20614
31. S. Baroni et al., *Rev. Mod. Phys.* **73** (2), 515 (2001). (2001). doi:10.1103/RevModPhys.73.515
32. A. P. Bartok et al., *Phys. Rev. Lett.* **104**, 136403 (2010). doi:10.1103/PhysRevLett.104.136403
33. H. Babaei et al., *Phys. Rev. Mater.* **3** (7), 074603 (2019). doi:10.1103/PhysRevMaterials.3.074603
34. D. Dragoni et al., *Phys. Rev. Mater.* **2** (1), 031808 (2018). doi:10.1103/PhysRevMaterials.2.013808
35. Z. Chen et al., *Adv. Sci.* 8 (12), 2004214 (2021). doi:10.1002/advs.202004214
36. A. M. Samarakoon et al., *Nat. Commun.* **11** (1), 892 (2020). doi:10.1038/s41467-020-14660-y
37. J. Liu et al., *Phys. Rev. B.* **95** (4), 041101(R) (2017). doi:10.1103/PhysRevB.95.041101
38. L. Huang and L. Wang, *Phys. Rev. B.* **95** (3), 035105 (2017). doi:10.1103/PhysRevB.95.035105
39. J. Bonca et al., *Open Problems in Strongly Correlated Electron Systems* (Springer, Amsterdam, 2001).
40. M. Mitrano and Y. Wang, *Commun. Phys.* **3** (1), 184 (2020). doi:10.1038/s42005-020-00447-6
41. Y. Wang et al., *Nat. Rev. Mater.* **3** (9), 312 (2018). doi:10.1038/s41578-018-0046-3
42. Y. Chen et al., *Phys. Rev. B.* **99** (10), 104306 (2019). doi:10.1103/PhysRevB.99.104306
43. D. Christiansen et al., *Phys. Rev. B.* **100** (20), 205401 (2019). doi:10.1103/PhysRevB.100.205401