Acta Cryst A FOUNDATIONS ADVANCES

ISSN 2053-2733

Machine learning in crystallography and structural science

Simon J. L. Billinge[a]* and Thomas Proffen[b]*

[a]Department of Applied Physics and Applied Mathematics, Columbia University, New York, USA, and [b]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA. *Correspondence e-mail: sb2896@columbia.edu, tproffen@ornl.gov

Keywords: machine learning; artificial intelligence; deep learning; neural networks; IUCr Journals.

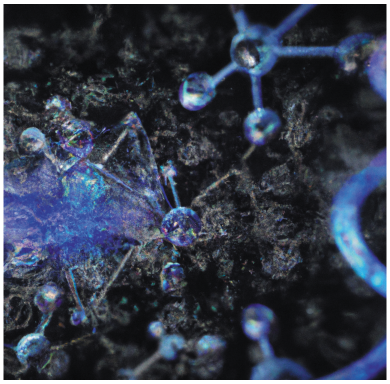
We are happy to present a virtual collection of articles from the journals of the International Union of Crystallography (IUCr) dealing with the application of artificial intelligence (AI) and machine learning (ML) in structural science (https://journals.iucr.org/special_issues/2024/ML/). AI/ML is revolutionizing our everyday lives.

Although the foundations of machine learning and deep learning (DL) came from the worlds of academic computing, mathematics and theories of the brain (McCulloch & Pitts, 1943; Rosenblatt, 1958), many of the early societal impacts were in commerce. However, physical scientists are now adopting these developments in the pursuit of their own science (Choudhary *et al.*, 2021), and crystallography is no exception. It is therefore very timely to pull together the growing number of AI/ML papers that have been published in *Acta Crystallographica* (*Sections A*, *B* and *D*), *IUCrJ* and *Journal of Synchrotron Radiation*. We also note a related virtual collection on AI published in the *Journal of Applied Crystallography* at https://journals.iucr.org/special_issues/2024/ANNs/ and the recent lead article in *Acta Crystallographica Section A* on deep learning applications in protein crystallography (Matinyan *et al.*, 2024).

The purpose of this article is not to review each of the papers in the virtual collection, but instead to encourage you to explore the papers in their own right. In Table 1 we have therefore summarized both the scientific target and the AI/ML method used in each paper, allowing you to quickly navigate to papers of greatest interest to you. In this article we seek to provide some higher-level themes and group some of the papers by ML and domain topics in an attempt to help you gain an appreciation of how the field has developed in crystallography and how scientists are currently using AI/ML as a tool to solve their scientific problems.

Virtually all of the types of ML are represented among these papers. Unsupervised learning is an approach where ML algorithms are shown sets of data with no prior knowledge and they attempt to cluster them (*i.e.* find similar signals) or extract reduced sets of distinct signals that can explain the behavior of a larger set of signals. In supervised learning, algorithms are 'trained' on large sets of prior data, after which they can classify new data that they are given based on what they learned from the training data. This classification problem is exemplified by training algorithms to differentiate between pictures of cats and dogs (Subramanian, 2018). Supervised learning can also be used to carry out regression rather than classification, carrying out function fitting to sets of data. Finally, various generative ML approaches aim to generate new outputs given some input prompts that are based on training on large amounts of learned responses. Deepfake video and audio technologies and ChatGPT (OpenAI, 2024*a*) are examples of generative AI.

Another approach for differentiating different AI/ML approaches is based on the internal structure of the algorithm. Broadly speaking, these can be divided into conventional ML and deep neural nets (deep learning, DL, for short). The conventional methods are based on statistical methods and linear algebra, and include tree-based methods, logistic regression and matrix factorization approaches. In deep learning, highly nonlinear graphical mathematical structures are constructed, inspired by the neuron structures of the brain, with information being passed through the network from an input side to an output side, whilst undergoing nonlinear transformations at each level. The transformations and passage of data through the networks are controlled by many thousands of parameters that are algorithmically updated to allow the network to make



https://doi.org/10.1107/S2053273324000172

accurate mappings of various known inputs to their known outputs. This is the training stage. Once trained, new inputs that the network has not seen before are given to the network and it predicts the output, which is compared with the known output; these are the validating and testing stages. The training, validation and testing stages are iterated until the network results in satisfactory predictive power, at which point it may be put into production so that it makes predictions from inputs with unknown outputs. Deep neural nets tend to make better predictions than conventional ML approaches and are often preferred in production, at the expense of needing more training data, requiring more computing power and having behavior that is less intelligible to the operator.

The earliest AI publication in the IUCr journals is, rather remarkably, from 1977 (Feigenbaum *et al.*, 1977), a prior epoch of AM/ML, where a proposal is made to apply AI to protein crystallography. The authors tackled the problem of assigning amino-acid sequences to electron-density maps, mapping it onto a classical ‘scene analysis’ problem in robotics and computer vision – the ‘blocks world’ problem where a robot is tasked with recreating the 3D scene from a blurry 2D television image so as to manipulate 3D wooden blocks. This topic was picked up again in the early 1990s in *Acta Crystallographica Section D* in an attempt to incorporate prior structural information into direct-methods approaches for protein structure solution, extending scene analysis to ‘molecular scenes’ (Fortier *et al.*, 1993).

The next AI papers did not appear in *Acta Crystallographica* until 2002 (Christensen, 2002; Ioerger & Sacchettini, 2002), a full 25 years after the first, but still a solid 10–15 years before the golden days of the latest ML epoch. Both papers describe the use of a feed-forward neural net, or multilayer perceptron (MLP), with an input and an output layer but only one hidden layer (Fig. 1) – a predecessor to latter-day deep neural nets.

Christiansen (2002) used it to predict which type of atom sits within each Voronoi polyhedron computed from the coordinates of the atoms in a crystal structure. The MLP was trained as a binary classifier that would predict whether each polyhedron in the tessellation contained C or H from four input quantities related to the geometry of the Voronoi polyhedron. The goal was somewhat modest, but it was shown to work, being trained from data held in structural databases. Ioerger & Sacchettini (2002) used their MLP to try to automate the procedure of assigning C$^{\alpha}$ atoms in a protein to peaks in the electron density that had previously been determined by direct methods.

During this period a number of papers appeared addressing the problem of protein crystallization; not crystallography directly, but a major bottleneck in protein structure solution at the time (Berntson *et al.*, 2003; Gopalakrishnan *et al.*, 2004; Liu *et al.*, 2008; Jahandideh *et al.*, 2014). Gopalakrishnan *et al.* (2004) used the Biological Macromolecule Crystallization Database (BMCD) (Gilliland, 1988) with modest success to predict synthesis conditions conducive to protein crystallization, still an unsolved problem. The challenge was the paucity of data, and rule learning algorithms were tried as early attempts at feature engineering and incorporation of domain knowledge in the ML approach. AI-enabled high-throughput screening of diffraction images was also explored (Berntson *et al.*, 2003) as an exploratory exercise using novel shallow neural nets called correlation cascade nets.

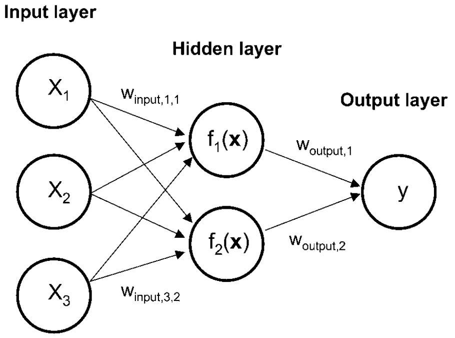


Figure 1
The multi-level perceptron, an early, shallow, neural net, reported in Christensen (2002).

ML reappeared in *Acta A* in 2016 (Muthig *et al.*, 2016), a further 14 years after the previous AI paper, where statistical approaches to carrying out the inverse Fourier transform to obtain $P(r)$ from small-angle-scattering data were tested, and the results post-processed using ML to remove ripple artifacts. Cross validation, an approach of ML, was used to determine the crossover from underfitting to overfitting with increasing model complexity, and the relevance vector machine (RVM) and least absolute shrinkage and selection operator (LASSO) conventional ML approaches were used to improve model stability.

Starting in 2017 (Park *et al.*, 2017), with a deep convolutional neural net used to classify the crystal system and space group of simulated powder diffraction patterns, an explosion of activity followed in 2019 and the modern period of ML applied to crystallography fully started, with five AI/ML papers appearing in *Acta A* alone in that year (Conterosito *et al.*, 2019; Gao *et al.*, 2019; Liu *et al.*, 2019; Garcia-Bonete & Kantona, 2019; Song *et al.*, 2019). These ranged from the use of principal component analysis (PCA), an unsupervised machine-learning approach applied to the study of CO_2 adsorption in zeolite-Y (Conterosito *et al.*, 2019), to applying convolutional neural nets to predict the space group of a structure given just its atomic pair distribution function (PDF) as input (Liu *et al.*, 2019). The latter model is now in production as the *spacegroupMining* (Yang *et al.*, 2021) web service at https://pdfitc.org, as an example of how trained ML models may be deployed to help the community in their everyday scientific endeavors.

The advances in deep learning from 2003 to 2019 are profound, taking us from a network with a single hidden layer binary classifier that chose C or H for each Voronoi polyhedron to the deep neural net in Liu *et al.* (2019), which could successfully classify experimental PDFs into 45 space groups with >90% top-six accuracy with only the PDF signal itself as input (after being trained on ~80 000 known structures).

Table 1
The crystallographic topic and machine-learning approach of papers in the virtual collection.

The classification of the crystallographic problems and machine-learning approaches in this table is intended to help the readers and provide a rough guide. In some cases, it might over-simplify the approach taken and/or problem solved. Abbreviations: ML: machine learning; NN: neural network; DL: deep learning; RVM: relevance vector machine; LASSO: least absolute shrinkage and selection operator; CT: computed tomography; cryo-EM: cryo-electron microscopy; DFT: density functional theory; EM: electron microscopy; PDF: pair distribution function; SAD: single-wavelength anomalous diffraction; SAS: small-angle scattering; XFEL: X-ray free-electron laser.

Year	Crystallographic problem	Machine-learning approach	Citation
Acta Crystallographica Section A			
1977	Assigning amino-acid sequences to electron-density maps	Image recognition, scene analysis	Feigenbaum, E. A., Engelmore, R. S. & Johnson, C. K. (1977). *Acta Cryst.* A**33**, 13–18.
2002	Binary classifier to predict the type of nearest neighbor atoms between C or H. Pioneering use of neural net	Supervised ML, classification, feed-forward (shallow) neural net	Christensen, S. W. (2002). *Acta Cryst.* A**58**, 171–179.
2016	Obtaining the $P(r)$ correlation function from small-angle-scattering data	Inverse Fourier transform using statistical inference and the help of RVM and LASSO ML methods, regression	Muthig, M., Prévost, S., Orglmeister, R. & Gradzielski, M. (2016). *Acta Cryst.* A**72**, 557–569.
2019	CO_2 adsorption into zeolite-Y	Unsupervised ML, principal component analysis	Conterosito, E., Palin, L., Caliandro, R., van Beek, W., Chernyshov, D. & Milanesio, M. (2019). *Acta Cryst.* A**75**, 214–222.
2019	Classification of atomic PDF data by space group	Supervised ML, convolutional DL	Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. (2019). *Acta Cryst.* A**75**, 633–643.
2019	Improved phases from SAD data	Multivariate Bayesian analysis, regression	Garcia-Bonete, M.-J. & Katona, G. (2019). *Acta Cryst.* A**75**, 851–860.
2019	Indexing of synchrotron Laue X-ray microdiffraction scans	Convolutional NN autoencoder, supervised ML, classification, image segmentation	Song, Y., Tamura, N., Zhang, C., Karami, M. & Chen, X. (2019). *Acta Cryst.* A**75**, 876–888.
2020	Structure determination from PDF	Model selection, database screening	Yang, L., Juhás, P., Terban, M. W., Tucker, M. G. & Billinge, S. J. L. (2020). *Acta Cryst.* A**76**, 395–409.
2021	Pair distribution function analysis	Database infrastructure	Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjær, E. T. S., Jensen, K. M. Ø., Tucker, M. G. & Billinge, S. J. L. (2021). *Acta Cryst.* A**77**, 2–6.
2022	Literature search via powder data	Unsupervised ML	Özer, B., Karlsen, M. A., Thatcher, Z., Lan, L., McMahon, B., Strickland, P. R., Westrip, S. P., Sang, K. S., Billing, D. G., Ravnsbæk, D. B. & Billinge, S. J. L. (2022). *Acta Cryst.* A**78**, 386–394.
2023	Structure modeling, force fields for DFT	Supervised ML, classification	Hofmann, D. W. M. & Kuleshova, L. N. (2023). *Acta Cryst.* A**79**, 132–144.
2023	Optical crystallization screening	Supervised ML, classification, convolutional NNs	Thielmann, Y., Luft, T., Zint, N. & Koepke, J. (2023). *Acta Cryst.* A**79**, 331–338.
2023	Cryo-EM data selection	Supervised ML, DL, classification	Matinyan, S., Demir, B., Filipcik, P., Abrahams, J. P. & van Genderen, E. (2023). *Acta Cryst.* A**79**, 360–368.
2024	Review of DL applications in protein crystallography	Various	Matinyan, S., Filipcik, P. & Abrahams, J. P. (2024). *Acta Cryst.* A**80**, 1–17.
Acta Crystallographica Section B			
2015	Perovskite classification	Unsupervised ML	Pilania, G., Balachandran, P. V., Gubernatis, J. E. & Lookman, T. (2015). *Acta Cryst.* B**71**, 507–513.
2017	Predicting structural displacements	Supervised ML (from DFT calculations)	Balachandran, P. V., Shearman, T., Theiler, J. & Lookman, T. (2017). *Acta Cryst.* B**73**, 962–967.
Acta Crystallographica Section D			
1993	Solving protein structures, data-augmented direct methods	Knowledge representations, scene analysis	Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Acta Cryst.* D**49**, 168–178.
2002	Extracting protein structure from electron-density maps. Pioneering use of neural net	Supervised ML, regression, feed-forward (shallow) neural net	Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.
2004	Predicting criteria for protein crystallization	Supervised ML, rule learner approaches	Gopalakrishnan, V., Livingston, G., Hennessy, D., Buchanan, B. & Rosenberg, J. M. (2004). *Acta Cryst.* D**60**, 1705–1716.
2008	Protein crystal detection/screening	Unsupervised ML, image recognition, decision trees	Liu, R., Freund, Y. & Spraggon, G. (2008). *Acta Cryst.* D**64**, 1187–1195.
2014	Predicting factors that affect protein crystallization	Supervised ML, random forest	Jahandideh, S., Jaroszewski, L. & Godzik, A. (2014). *Acta Cryst.* D**70**, 627–635.
2019	Protein residue classification	Supervised ML, classification	Chojnowski, G., Pereira, J. & Lamzin, V. S. (2019). *Acta Cryst.* D**75**, 753–763.
2020	Protein model correctness determination	Supervised ML, classification	Bond, P. S., Wilson, K. S. & Cowtan, K. D. (2020). *Acta Cryst.* D**76**, 713–723.

Table 1 (continued)

Year	Crystallographic problem	Machine-learning approach	Citation
2021	Structure determination, phasing	Unsupervised ML	McCoy, A. J., Stockwell, D. H., Sammito, M. D., Oeffner, R. D., Hatti, K. S., Croll, T. I. & Read, R. J. (2021). *Acta Cryst.* D**77**, 1–10.
2021	Protein structure prediction	*AlphaFold*	Bouatta, N., Sorger, P. & AlQuraishi, M. (2021). *Acta Cryst.* D**77**, 982–991.
2021	Protein structure solution	*AlphaFold*	Moroz, O. V., Blagova, E., Lebedev, A. A., Sánchez Rodríguez, F., Rigden, D. J., Tams, J. W., Wilting, R., Vester, J. K., Longhin, E., Hansen, G. H., Krogh, K. B. R. M., Pache, R. A., Davies, G. J. & Wilson, K. S. (2021). *Acta Cryst.* D**77**, 1564–1578.
2022	Protein structure determination	*AlphaFold*	McCoy, A. J., Sammito, M. D. & Read, R. J. (2022). *Acta Cryst.* D**78**, 1–13.
2022	Diffraction artifact removal (here from ice)	Convolutional NN, supervised ML	Nolte, K., Gao, Y., Stäb, S., Kollmannsberger, P. & Thorn, A. (2022). *Acta Cryst.* D**78**, 187–195.
2022	Protein structure prediction	DL, *AlphaFold*	Barbarin-Bocahu, I. & Graille, M. (2022). *Acta Cryst.* D**78**, 517–531.
2022	Cryo-EM data cleaning (particle pruning)	DL, convolutional NN, supervised ML	Sánchez Rodríguez, F., Chojnowski, G., Keegan, R. M. & Rigden, D. J. (2022). *Acta Cryst.* D**78**, 1412–1427.
2023	Protein structure determination	*AlphaFold*	Terashi, G., Wang, X. & Kihara, D. (2023). *Acta Cryst.* D**79**, 10–21.
2023	Protein structure determination	*AlphaFold*	Terwilliger, T. C., Afonine, P. V., Liebschner, D., Croll, T. I., McCoy, A. J., Oeffner, R. D., Williams, C. J., Poon, B. K., Richardson, J. S., Read, R. J. & Adams, P. D. (2023). *Acta Cryst.* D**79**, 234–244.
2023	Protein model building	Supervised ML	Alharbi, E., Calinescu, R. & Cowtan, K. (2023). *Acta Cryst.* D**79**, 326–338.
2023	Correcting systematic errors in protein diffraction data	Review and best practices for the use of Bayesian variational inference for correcting systematic errors in diffraction structure factors	Aldama, L. A., Dalton, K. M. & Hekstra, D. R. (2023). *Acta Cryst.* D**79**, 796–805.
2024	Simulation and characterization of protein diffraction images	Supervised ML, DL	Mendez, D., Holton, J. M., Lyubimov, A. Y., Hollatz, S., Mathews, I. I., Cichosz, A., Martirosyan, V., Zeng, T., Stofer, R., Liu, R., Song, J., McPhillips, S., Soltis, M. & Cohen, A. E. (2024). *Acta Cryst.* D**80**, 26–43.
Journal of Synchrotron Radiation			
2003	Protein crystal screening, high-throughput quality assessment of diffraction patterns from protein crystallization wells. Pioneering use of shallow NNs	Supervised ML, classification	Berntson, A., Stojanoff, V. & Takai, H. (2003). *J. Synchrotron Rad.* **10**, 445–449.
2010	Analysis of nuclear resonant scattering	Supervised ML	Planckaert, N., Demeulemeester, J., Laenens, B., Smeets, D., Meersschaut, J., L'abbé, C., Temst, K. & Vantomme, A. (2010). *J. Synchrotron Rad.* **17**, 86–92.
2017	CT reconstruction	Convolutional NN	Yang, X., De Carlo, F., Phatak, C. & Gürsoy, D. (2017). *J. Synchrotron Rad.* **24**, 469–475.
2018	Macromolecular crystal screen, Bragg-spot detection	Convolutional NN, supervised ML	Ke, T.-W., Brewster, A. S., Yu, S. X., Ushizima, D., Yang, C. & Sauter, N. K. (2018). *J. Synchrotron Rad.* **25**, 655–670.
2019	CT reconstruction, upsampling	Convolutional NNs, supervised ML	Bellos, D., Basham, M., Pridmore, T. & French, A. P. (2019). *J. Synchrotron Rad.* **26**, 839–853.
2019	Spectrometer design	Support vector machines, supervised ML	Li, Z. & Li, B. (2019). *J. Synchrotron Rad.* **26**, 1058–1068.
2019	Protein crystal centering	Convolutional NN, DL, supervised ML	Ito, S., Ueno, G. & Yamamoto, M. (2019). *J. Synchrotron Rad.* **26**, 1361–1366.
2020	CT reconstruction	Supervised ML, DL	Huang, Y., Wang, S., Guan, Y. & Maier, A. (2020). *J. Synchrotron Rad.* **27**, 477–485.
2020	CT calibration (rotation axis)	Convolutional NN, supervised ML	Yang, X., Kahnt, M., Brückner, D., Schropp, A., Fam, Y., Becher, J., Grunwaldt, J.-D., Sheppard, T. L. & Schroer, C. G. (2020). *J. Synchrotron Rad.* **27**, 486–493.
2020	Grazing-incidence SAS classification	Supervised ML, DL, convolutional NN, classification	Ikemoto, H., Yamamoto, K., Touyama, H., Yamashita, D., Nakamura, M. & Okuda, H. (2020). *J. Synchrotron Rad.* **27**, 1069–1073.
2020	Image filtering nanotomography	Supervised ML	Flenner, S., Storm, M., Kubec, A., Longo, E., Döring, F., Pelt, D. M., David, C., Müller, M. & Greving, I. (2020). *J. Synchrotron Rad.* **27**, 1339–1346.

Table 1 (continued)

Year	Crystallographic problem	Machine-learning approach	Citation
2021	CT image segmentation	DL, supervised ML	Ali, S., Mayo, S., Gostar, A. K., Tennakoon, R., Bab-Hadiashar, A., MCann, T., Tuhumury, H. & Favaro, J. (2021). *J. Synchrotron Rad.* **28**, 566–575.
2021	Signal processing, pulse shaping of synchrotron pulses	Convolutional NN, DL, supervised ML	Ma, X.-K., Huang, H.-Q., Ji, X., Dai, H.-Y., Wu, J.-H., Zhao, J., Yang, F., Tang, L., Jiang, K.-M., Ding, W.-C. & Zhou, W. (2021). *J. Synchrotron Rad.* **28**, 910–918.
2021	CT image correction	Supervised ML, transfer learning, DL	Fu, T., Zhang, K., Wang, Y., Li, J., Zhang, J., Yao, C., He, Q., Wang, S., Huang, W., Yuan, Q., Pianetta, P. & Liu, Y. (2021). *J. Synchrotron Rad.* **28**, 1909–1915.
2022	Image denoising, CT	Convolutional NN, self-supervised learning	Flenner, S., Bruns, S., Longo, E., Parnell, A. J., Stockhausen, K. E., Müller, M. & Greving, I. (2022). *J. Synchrotron Rad.* **29**, 230–238.
2022	CT segmentation	Supervised ML, DL	Gaudez, S., Ben Haj Slama, M., Kaestner, A. & Upadhyay, M. V. (2022). *J. Synchrotron Rad.* **29**, 1232–1240.
2022	X-ray emission analysis	Unsupervised ML	Hwang, I.-H., Solovyev, M. A., Han, S.-W., Chan, M. K. Y., Hammonds, J. P., Heald, S. M., Kelly, S. D., Schwarz, N., Zhang, X. & Sun, C.-J. (2022). *J. Synchrotron Rad.* **29**, 1309–1317.
2022	Digital twin model of a synchrotron undulator	Supervised ML, regression	Sheppard, R., Baribeau, C., Pedersen, T., Boland, M. & Bertwistle, D. (2022). *J. Synchrotron Rad.* **29**, 1368–1375.
2022	Region of interest finder, X-ray fluorescence microscopy	Unsupervised and supervised ML	Chowdhury, M. A. Z., Ok, K., Luo, Y., Liu, Z., Chen, S., O'Halloran, T. V., Kettimuthu, R. & Tekawade, A. (2022). *J. Synchrotron Rad.* **29**, 1495–1503.
2023	X-ray optics control	Supervised ML	Gunjala, G., Wojdyla, A., Goldberg, K. A., Qiao, Z., Shi, X., Assoufid, L. & Waller, L. (2023). *J. Synchrotron Rad.* **30**, 57–64.
2023	Diffraction data artifact detection	DL, convolutional NN, supervised ML	Yanxon, H., Weng, J., Parraga, H., Xu, W., Ruett, U. & Schwarz, N. (2023). *J. Synchrotron Rad.* **30**, 137–146.
2023	CT reconstruction	Supervised ML, DL	Fu, T., Wang, Y., Zhang, K., Zhang, J., Wang, S., Huang, W., Wang, Y., Yao, C., Zhou, C. & Yuan, Q. (2023). *J. Synchrotron Rad.* **30**, 620–626.
2023	CT imaging, dynamics, porous media	DL, supervised and unsupervised ML	Fokin, M. I., Nikitin, V. V. & Duchkov, A. A. (2023). *J. Synchrotron Rad.* **30**, 978–988.
2023	Reflectometry analysis, automation	Convolutional NN, supervised ML	Pithan, L., Starostin, V., Mareček, D., Petersdorf, L., Völter, C., Munteanu, V., Jankowski, M., Konovalov, O., Gerlach, A., Hinderhofer, A., Murphy, B., Kowarik, S. & Schreiber, F. (2023). *J. Synchrotron Rad.* **30**, 1064–1075.
2023	CT reconstruction	Convolutional NN	Cheng, C.-C., Chiang, M.-H., Yeh, C.-H., Lee, T.-T., Ching, Y.-T., Hwu, Y. & Chiang, A.-S. (2023). *J. Synchrotron Rad.* **30**, 1135–1142.
IUCrJ			
2017	Structure determination from powder diffraction (crystal system, extinction, space group)	DL, convolutional NN, supervised ML, classification	Park, W. B., Chung, J., Jung, J., Sohn, K., Singh, S. P., Pyo, M., Shin, N. & Sohn, K.-S. (2017). *IUCrJ*, **4**, 486–494.
2018	Particle pruning in cryo-EM images for single-particle structure determination	DL, supervised ML	Sanchez-Garcia, R., Segura, J., Maluenda, D., Carazo, J. M. & Sorzano, C. O. S. (2018). *IUCrJ*, **5**, 854–865.
2019	XFEL image classification	DL, convolutional NN, supervised ML	Shi, Y., Yin, K., Tai, X., DeMirci, H., Hosseinizadeh, A., Hogue, B. G., Li, H., Ourmazd, A., Schwander, P., Vartanyants, I. A., Yoon, C. H., Aquila, A. & Liu, H. (2019). *IUCrJ*, **6**, 331–340.
2019	EM image analysis	DL, supervised ML	Ramírez-Aportela, E., Mota, J., Conesa, P., Carazo, J. M. & Sorzano, C. O. S. (2019). *IUCrJ*, **6**, 1054–1063.
2020	Crystal picking in cryo-EM	Self-supervised ML, convolutional NN	McSweeney, D. M., McSweeney, S. M. & Liu, Q. (2020). *IUCrJ*, **7**, 719–727.
2020	Protein structure determination	DL, convolutional NN	Farrell, D. P., Anishchenko, I., Shakeel, S., Lauko, A., Passmore, L. A., Baker, D. & DiMaio, F. (2020). *IUCrJ*, **7**, 881–892.
2020	Structure stability prediction	Supervised and unsupervised ML, explainable ML	Pham, T.-L., Nguyen, D.-N., Ha, M.-Q., Kino, H., Miyake, T. & Dam, H.-C. (2020). *IUCrJ*, **7**, 1036–1047.

Table 1 (continued)

Year	Crystallographic problem	Machine-learning approach	Citation
2021	Phase domain imaging	DL, convolutional NN, supervised ML	Wu, L., Juhas, P., Yoo, S. & Robinson, I. (2021). *IUCrJ*, **8**, 12–21.
2021	Protein structure determination	Convolutional NN, DL, supervised ML	Kimanius, D., Zickert, G., Nakane, T., Adler, J., Lunz, S., Schönlieb, C.-B., Öktem, O. & Scheres, S. H. W. (2021). *IUCrJ*, **8**, 60–75.
2021	Phase identification from powder diffraction	DL, convolutional NN, supervised ML	Schuetzke, J., Benedix, A., Mikut, R. & Reischl, M. (2021). *IUCrJ*, **8**, 408–420.
2021	3D grain mapping	DL, supervised ML	Fang, H., Hovad, E., Zhang, Y., Clemmensen, L. K. H., Ersbøll, B. K. & Juul Jensen, D. (2021). *IUCrJ*, **8**, 719–731.
2022	Protein residue classification	Supervised ML, classification	Chojnowski, G., Simpkin, A. J., Leonardo, D. A., Seifert-Davila, W., Vivas-Ruiz, D. E., Keegan, R. M. & Rigden, D. J. (2022). *IUCrJ*, **9**, 86–97.
2022	Bragg-peak position determination	DL, supervised ML	Liu, Z., Sharma, H., Park, J.-S., Kenesei, P., Miceli, A., Almer, J., Kettimuthu, R. & Foster, I. (2022). *IUCrJ*, **9**, 104–113.
2022	Coherent diffraction feature extraction	Unsupervised ML	Pan, D., Fan, J., Nie, Z., Sun, Z., Zhang, J., Tong, Y., He, B., Song, C., Kohmura, Y., Yabashi, M., Ishikawa, T., Shen, Y. & Jiang, H. (2022). *IUCrJ*, **9**, 223–230.
2022	Cryo-EM image enhancement	DL, supervised ML	Ramírez-Aportela, E., Carazo, J. M. & Sorzano, C. O. S. (2022). *IUCrJ*, **9**, 632–638.
2023	RNA structure characterization	Supervised ML, classification, regression	Cheng, A., Kim, P. T., Kuang, H., Mendez, J. H., Chua, E. Y. D., Maruthi, K., Wei, H., Sawh, A., Aragon, M. F., Serbynovskyi, V., Neselu, K., Eng, E. T., Potter, C. S., Carragher, B., Bepler, T. & Noble, A. J. (2023). *IUCrJ*, **10**, 77–89.
2023	Cryo-EM automation	Convolutional NN, supervised ML	Kim, P. T., Noble, A. J., Cheng, A. & Bepler, T. (2023). *IUCrJ*, **10**, 90–102.
2023	Nanofiber orientation determination	DL, convolutional NN, supervised ML	Sun, M., Dong, Z., Wu, L., Yao, H., Niu, W., Xu, D., Chen, P., Gupta, H. S., Zhang, Y., Dong, Y., Chen, C. & Zhao, L. (2023). *IUCrJ*, **10**, 297–308.
2023	Protein structure solution	DL, convolutional NN, supervised ML	Pan, T., Jin, S., Miller, M. D., Kyrillidis, A. & Phillips, G. N. (2023). *IUCrJ*, **10**, 487–496.
2023	Unit cell from PDF	Supervised ML, DL, classification	Guccione, P., Diacono, D., Toso, S. & Caliandro, R. (2023). *IUCrJ*, **10**, 610–623.

Many of these early AI efforts were not highly successful and garnered few citations, but the impact of AI developments and applications in the structural science domain are only now being felt. The huge changes from the early 2000s to now are the availability of high-performance computing and a much greater abundance of training data. This illustrates a theme, in that much of the AI/ML used in crystallography and materials science is possible as a result of the availability of large databases of structures, which are in existence because of the early adoption of informatics approaches by crystallographers in the form of data standards for structures (*e.g.*, CIF and PDB) and the resulting structured databases (Groom *et al.*, 2016; Gates-Rector & Blanton, 2019; Levin, 2018; Gražulis *et al.*, 2009; Jain *et al.*, 2013; Berman *et al.*, 2000; Kirklin *et al.*, 2015), guided by commissions of the IUCr and encouraged, and later enforced, by its journals. Crystallography has been at the forefront of data analytics applied to materials science and structural biology and, as this collection indicates, remains so today.

Note: The image on the first page of this Editorial was chosen for the 'cover' of the virtual collection from a range of images generated by DALL·E (OpenAI, 2024*b*) using the prompt 'A depiction of molecules surrounded by abstract representations of digital data and AI algorithms, highlighting the historical improvements in the data-driven approach to crystallography'.

Funding information

SJLB acknowledges support from the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences (DOE-BES) under contract No. DE-SC0024141. Work at Oak Ridge National Laboratory was sponsored by the Scientific User Facilities Division, Office of Basic Energy Sciences, US Department of Energy.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Berntson, A., Stojanoff, V. & Takai, H. (2003). *J. Synchrotron Rad.* **10**, 445–449.

Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., WooPark, C., Choudhary, A., Agrawal, A., Billinge, S. J. L., Holm, E., Ong, S. P. & Wolverton, C. (2021). arXiv:2110.14820 [cond-Mater, physics: physics].

Christensen, S. W. (2002). *Acta Cryst.* A**58**, 171–179.

Conterosito, E., Palin, L., Caliandro, R., van Beek, W., Chernyshov, D. & Milanesio, M. (2019). *Acta Cryst.* A**75**, 214–222.

Feigenbaum, E. A., Engelmore, R. S. & Johnson, C. K. (1977). *Acta Cryst.* A**33**, 13–18.

Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Acta Cryst.* D**49**, 168–178.

Gao, Z., Guizar-Sicairos, M., Lutz-Bueno, V., Schröter, A., Liebi, M., Rudin, M. & Georgiadis, M. (2019). *Acta Cryst.* A**75**, 223–238.

Garcia-Bonete, M.-J. & Katona, G. (2019). *Acta Cryst.* A**75**, 851–860.

Gates-Rector, S. & Blanton, T. (2019). *Powder Diffr.* **34**, 352–360.

Gilliland, G. L. (1988). *J. Cryst. Growth*, **90**, 51–59.

Gopalakrishnan, V., Livingston, G., Hennessy, D., Buchanan, B. & Rosenberg, J. M. (2004). *Acta Cryst.* D**60**, 1705–1716.

Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. (2016). *Acta Cryst.* B**72**, 171–179.

Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.

Jahandideh, S., Jaroszewski, L. & Godzik, A. (2014). *Acta Cryst.* D**70**, 627–635.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. & Persson, K. A. (2013). *APL Mater.* **1**, 011002.

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S. & Wolverton, C. (2015). *npj Comput. Mater.* **1**, 15010.

Levin, I. (2018). NIST Inorganic Crystal Structure Database (ICSD), National Institute of Standards and Technology, https://doi.org/10.18434/M32147.

Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. (2019). *Acta Cryst.* A**75**, 633–643.

Liu, R., Freund, Y. & Spraggon, G. (2008). *Acta Cryst.* D**64**, 1187–1195.

Matinyan, S., Filipcik, P. & Abrahams, J. P. (2024). *Acta Cryst.* A**80**, 1–17.

McCulloch, W. S. & Pitts, W. (1943). *Bull. Math. Biophys.* **5**, 115–133.

Muthig, M., Prévost, S., Orglmeister, R. & Gradzielski, M. (2016). *Acta Cryst.* A**72**, 557–569.

OpenAI (2024*a*). ChatGPT. https://chat.openai.com.

OpenAI (2024*b*). DALL·E. https://openai.com/dall-e-3.

Park, W. B., Chung, J., Jung, J., Sohn, K., Singh, S. P., Pyo, M., Shin, N. & Sohn, K.-S. (2017). *IUCrJ*, **4**, 486–494.

Rosenblatt, F. (1958). *Psychol. Rev.* **65**, 386–408.

Song, Y., Tamura, N., Zhang, C., Karami, M. & Chen, X. (2019). *Acta Cryst.* A**75**, 876–888.

Subramanian, V. (2018). *Deep Learning with PyTorch: A Practical Approach to Building Neural Network Models Using PyTorch.* Packt Publishing Ltd.

Yang, L., Culbertson, E. A., Thomas, N. K., Vuong, H. T., Kjaer, E. T. S., Jensen, K. M. Ø., Tucker, M. G. & Billinge, S. J. L. (2021). *Acta Cryst.* A**77**, 2–6.