Session III, May 26, Tuesday

L6

THE USE OF MACHINE LEARNING INTERATOMIC POTENTIALS FOR CRYSTAL STRUCTURE SOLUTION VERIFICATION

M. Hušák

University of Chemistry and Technology, Prague, Technická 5, 166 28 Praha 6 – Dejvice
husakm@vscht.cz

Introduction

Correctly solved crystal structure should be in agreement with experimental data and its geometry should be located in a local minimum on a Potential Energy Surface (PES). The idea to verify crystal structure by comparison with DFT calculation results was introduced already 12 years ago [1,2]. Machine Learning Interatomic Potentials (MLIP) offer a significantly faster alternative to DFT as a tool for theoretical crystal structure calculation with comparable accuracy and correctness. The geometry optimized structures based on UMA(S) MLIP (parameterised on OMC25 dataset) [3] were used to verify 216 911 selected structures from Cambridge Structure Database (CSD). The targeted was to detect issues with both original X-ray experiments interpretation as well as with the MLIP functionality.

Methodology

Before the structure validation we had pre-filtered from CSD 6.0 a subset of structures, for which it had sense to run the test (See Table1). We had rejected structures with disorder, used MLIP non supported elements, voids, incorrect space group and to old structures measured on historical devices. Current MLIP do not support well electrostatic long-range forces so we had rejected structures with charged atoms as well. From used GPU memory limits reasons structures with unit cell volume more than 4000 Å^3 were not investigated. See Table 1 for the pre-filtering results.

The final MLIP calculation was running 4 months in parallel on 2 computers with NVIDIA RTX PRO 4000 Blackwell graphic card equipped by 24GB/ GDDR7 memory. Execution was done almost exclusively in GPU on 8 960 CUDA cores. The processing and descriptors calculation was driven by our proprietary software checkCIF-DFT.

The original verification method [1,2] was primary based on the experimental and DFT results comparison by Cartesian displacement (RMSCD) descriptor only. RMSCD is de-facto a RMSD modified to be able to compare atomic positions in different unit cells. It was already mentioned [1] this descriptor was not able to clearly separate e.g. totally artificial fraud structures from correctly solved one. We had tested several other descriptors to separate problematic structures from correct ones. Really prob-

Table 1. Result of structures pre-filtering

Rejection reason	Structures rejected based on given test	Structures left after test
Total structures in CSD 6.0 dataset		1371757
Presence of other elements than supported one (H,C,N,O,S,P,F,Cl,Br,I)	853820	517937
Disordered structure	90765	427172
Future not satisfied: No error, 3D coordinates present, single crystal data	32241	394931
Not published after 2004	95698	299233
Ionized molecules or charged atoms present	41768	257465
Volume over 4000 Å^3	17725	239740
3D coordinates not present (not covered by previous check)	4	239736
Solvent accessible volume over 40 Å^3	16463	223273
Incorrect space group or super cell detected	1128	222145
Incorrect element valence detected	3981	218164
Detected atoms marked by "?" indicating disorder	1253	216911

During examination of our results we had identified issues with multiple structures deposited in CSD. It should be noted that based on our finding the use of CSD as reliable source of molecular geometry information is questionable. Marking the problematic structures in the database with some other flack than “no error” will be suitable. On the other side we had identified issues with the used MLIP as well - incorrect proton transfer and incorrect geometry for

Table 2. Issues detected during the OSD type 1 structures test

Problem detection reason	Structures rejected based on given test	Structures left after test
Initial dataset		216911
Non existing (Unknown) element present	1	216910
CIF space group does not correspond to original CIF	1	216909
Bond breaking (not including H-atom)	466	216443
Bond breaking (H-atom involved)	1400	215043
RMSCD over 1.0 Å	167	214876
RMSCD over 0.25 Å	3167	211709

lematic structures can be detected efficiently by bond-breaking (see Table 2 and Figure 1).

Results

We had manually inspected 466 structures with bond breaking (hydrogen not involved) and 167 structures with RMSCD higher than 1.0 Å. The inspection was based on comparison with original X-ray data, higher level full DFT calculation and inspection of the original publications. We had identified several types of errors with the CSD deposited data: element type assignment miss-match, missing hydrogen, incorrect hydrogen placement or incorrect experimental information in the CIF (pressure typically). We had identified issues with the used MLIP PES generation as well - incorrect proton transfer and pure results for elements not often present in teaching dataset. Most of the PES related issues were related to the PBE+D3 functional used for OMC25 dataset generation. PBE-D3 shows known errors related to ion transfer. A better teaching set utilizing e.g. r2SCAN-MBD meta-GGA functional will be more suitable for future work. In rare situation we had detected as problematic correctly solved structures with non-usual futures. .

Conclusions

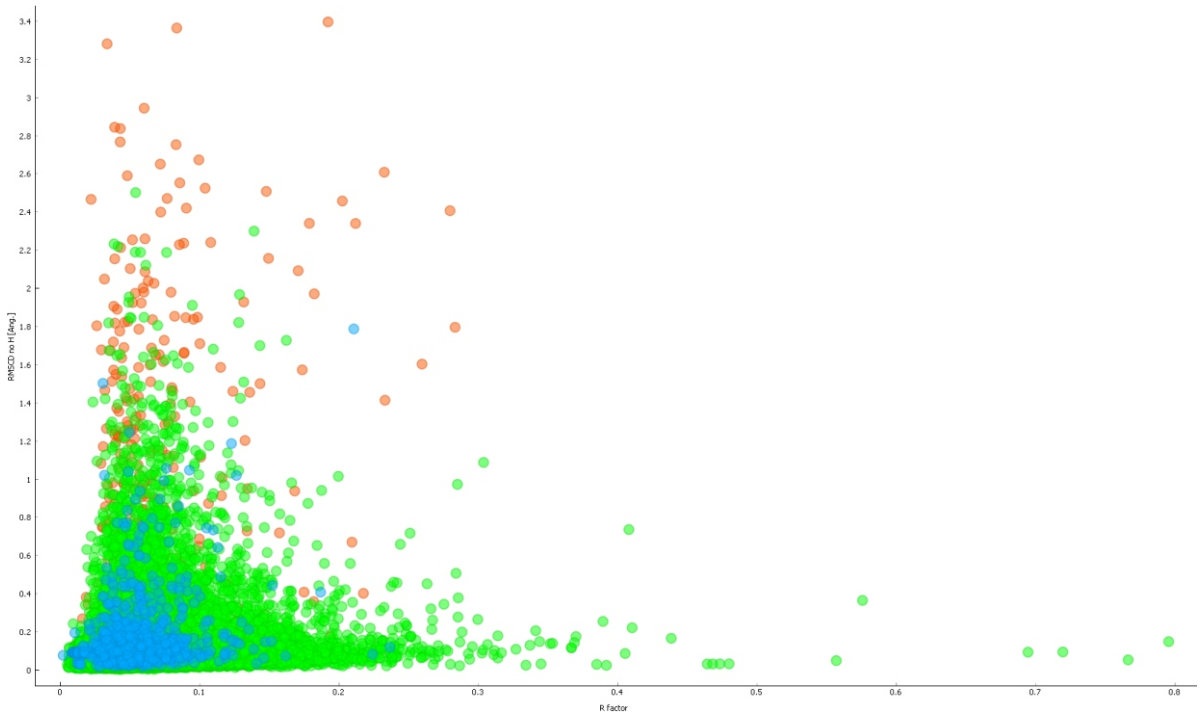


Figure 1. R factor versus RMSCD (hydrogen atoms excluded). green - no bond breaking, blue - only bonds including hydrogen break, red - breaking of bonds not including hydrogen.

computational time required for single structures sets is typically less then one minute - the described methodology is suitable for routine structure checks.

1. J. Streek, M. A. Neumann, *Acta .Cryst.*, **B66**, (2010), 544.
2. J. Streek, M. A. Neumann, *Acta .Cryst.*, **B70**, (2014), 1020.
3. Wood, B. M., Dzamba, M., Fu, X., Gao, M., Shuaibi, M., Barroso-Luque, L., Abdelmaqsoud, K., Gharakhanyan, V., Kitchin, J. R., Levine, D. S., Michel, K., Sriram, A., Cohen, T., Das, A., Rizvi, A., Sahoo, S. J., Ulissi, Z. W. & Zitnick, C. L. (2026). Uma: A family of universal models for atoms. https://doi.org/10.48550/arXiv.2506.23971.

L10

EXPLORING LOCAL ATOMIC ENVIRONMENTS USING UNIVERSAL MACHINE LEARNIG POTENTIALS AND AUTOMATED STRUCTURE GENERATION

J. Drahokoupil[1,2,3], M. Lebeda[1,2,3], Z. Holman[1,2]

[1]Fyzikální ústav, Akademie věd České Republiky, Na Slovance 2, 182 21 Praha 8, Česká Republika
[2]Katedra inženýrství pevných látek, Fakulta jaderná a fyzikálně inženýrská v Praze, České vysoké učení technické v Praze, Technická 4, 166 07 Praha 6 - Dejvice, Česká Republika
[3]Ústav fyziky, Fakulta strojní, České vysoké učení technické v Praze, Technická 4, 166 07 Praha 6 - Dejvice, Česká Republika
draho@fzu.cz

In the last few years, universal machine-learned interatomic potentials (UMLIPs) have entered the world of computer simulations, filling the gap between classical potentials and DFT calculations. Ideally, they achieve accuracy comparable to DFT but with a computational cost several orders of magnitude lower. Furthermore, their scaling is not cubic with respect to the number of atoms, making it possible to simulate structures containing hundreds of atoms even on a high-end desktop PC.

From the perspective of diffraction techniques, we usually observe an average structure where atomic positions may have partial occupancies. However, when describing structures for MD or DFT calculations, each atomic site must (with few exceptions) be either fully occupied or vacant. Therefore, if we want to simulate an alloy with a random atomic distribution, we must first generate such a configuration. Finding an arrangement that best mimics a truly random distribution is crucial for DFT calculations, where we typically work with small simulation cells. For this purpose, the concept of Special Quasi-random Structures (SQS) was introduced. While the **ATAT mcsqs** [1] program is commonly used to generate them, it isn't exactly user-friendly. To address this, our group developed **SimplySQS** [2], a tool designed to make this process much easier.

If the studied materials tend to exhibit atomic ordering, it is ideal to explore all possible configurations. This is practically feasible only for structures with a small number of atoms, as the number of combinations grows rapidly. A useful tool for generating various ordered structures is the **Supercell** [3] program, which also allows for the filtering of symmetry-equivalent structures.

To facilitate the use of universal potentials, our group is developing **uMLIP-interactive** [4]. This tool currently supports several universal ML potentials, including MACE, CHGNet, Nequix, SevenNet, Orb-v3, MatterSim, UPET, and GRACE. Through either a graphical interface or Python scripts, users can perform standard MD tasks such as energy calculation, geometry optimization, determination of elastic constants, phonon spectra calculations,

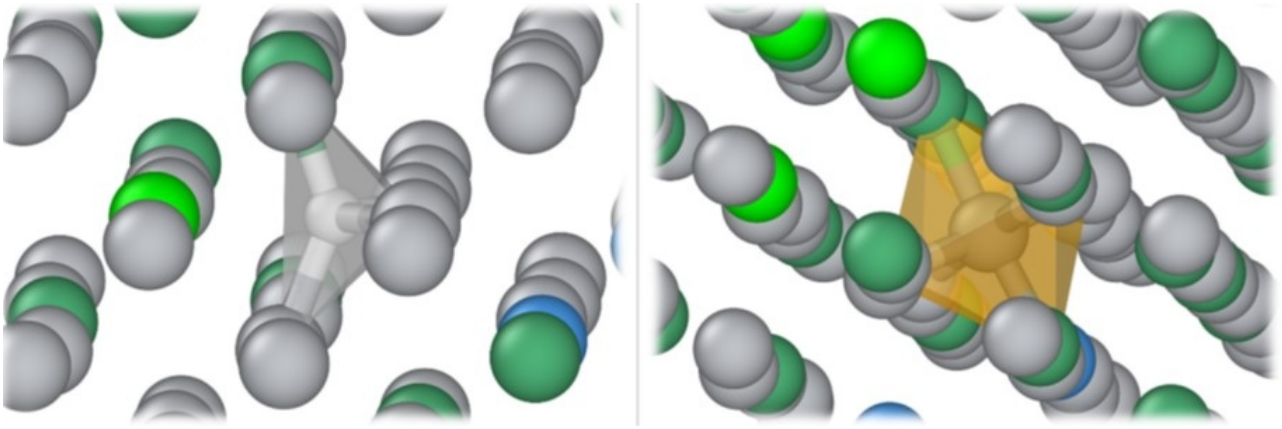


Figure 1. Stable tetrahedral position for a hydrogen atom (left) and stable octahedral position for a carbon atom (right) in the Ti-23Nb-0.7Ta-2Zr alloy.

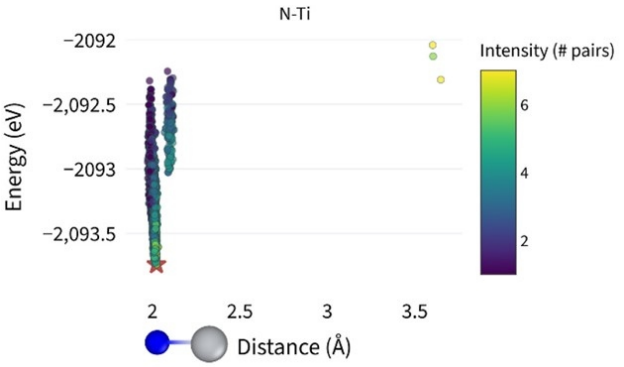


Figure 2. Energy dependence on the local environment of nitrogen atoms. It is evident that nitrogen atoms prefer environments dominated by titanium (Ti) atoms.

and various dynamical simulations (NVE, NVT, NPT). Input files are supported in cif, POSCAR, lmp, or xyz (with lattice info) formats.

In this contribution, we will discuss three specific examples of use:

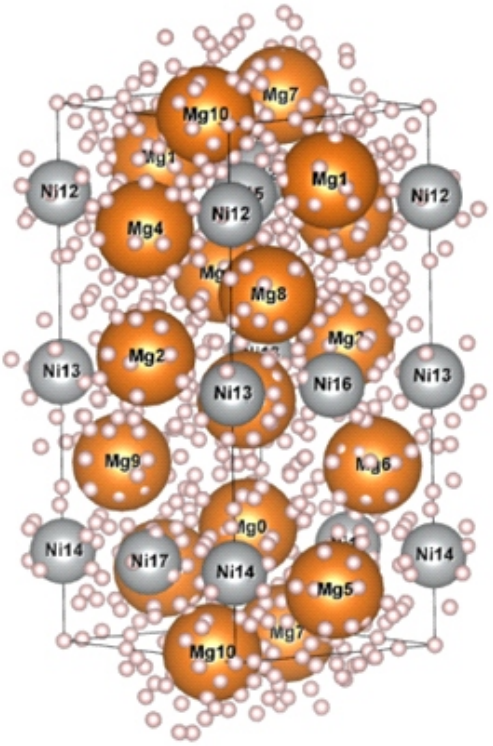
Preferential positions of gaseous elements in the "Gum Metal" alloy based on the Ti-23Nb-0.7Ta-2Zr composition.

When a small amount of oxygen (approx. 0.5%) is added to the Ti-23Nb-0.7Ta-2Zr alloy, it acquires remarkable properties: an elastic modulus close to 50 GPa combined with a strength exceeding 1000 MPa. Due to these unique characteristics, the alloy is referred to as "Gum Metal." Our research focused on identifying the preferential positions of interstitial atoms within this structure for low concentrations of gaseous elements: O, N, C, and H.

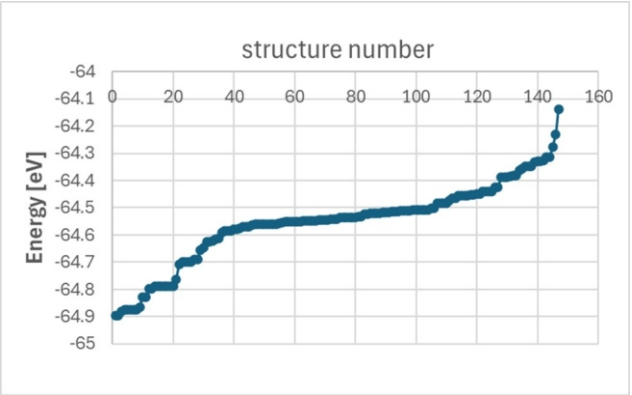
First, we transformed a simple bcc cell into a supercell containing 250 atoms and searched for an SQS structure corresponding to the composition in atomic percent (185 Ti, 58 Nb, 5 Zr, 2 Ta). From three independent runs, we obtained three distinct SQS structures. Within these structures, there are 750 octahedral and 1500 tetrahedral interstitial sites. We placed a single gas atom into each of these positions, performed geometry optimization, and statistically evaluated the entire data set. The results clearly show that O, N, and C prefer the larger octahedral sites, whereas H prefers tetrahedral sites. Furthermore, we found that for these interstitial atoms, a Ti-rich environment is energetically favorable, while an Nb-rich environment is unfavorable, see *Fig. 2*.

Hydrogen atom positions in Mg_2Ni.

The compound Mg_2Ni, sometimes enriched with rare earth elements, is a highly promising material for hydrogen storage, where the target stable phase is Mg_2NiH_4. Unfortunately, the saturation process can slow down in certain cases, making it difficult to load further hydrogen into the material. Understanding which sites hydrogen prefers within the original Mg_2Ni structure is of great interest, and not only from a diffusion perspective. For low hydrogen concentrations, diffraction techniques do not always provide a clear answer regarding the occupied positions.



a)



b)

Figure 3. a) The structure Mg_2Ni of with available interstitial sites for potential hydrogen localization (small pink spheres). b) Energies of individual structures containing two hydrogen atoms.

Therefore, we attempted to determine the hydrogen sites by calculating the energy for various potential locations. Using the **Point-Defect-Generator** [5], another tool being developed in our team, we performed a grid search to identify all potential hydrogen positions (see Fig. 3a). Subsequently, we created a structure for each possibility containing a single hydrogen atom and performed geometry optimization using a universal potential. We then placed an additional hydrogen atom into the lowest-energy structure, and so on.

Structure prediction of Mn – Hf – O.

The motivation for this research was the attempt to synthesize the $MnHfO_3$ phase, which, according to theoretical calculations, should exhibit interesting magnetic proper-

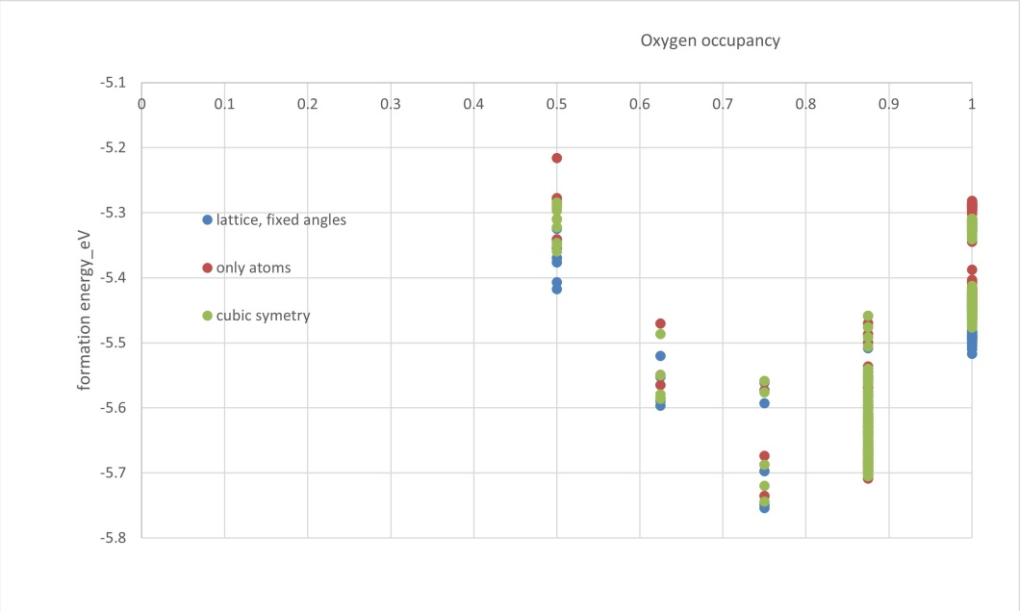


Figure 4. Formation energy of the Hf/MnO_2 structure as a function of oxygen vacancy concentration. Simulations were performed for various atomic arrangements within the cubic HfO_2 structure. The Supercell program was used to generate the individual configurations.

ties. Preparation in powder form was unsuccessful, resulting instead in separated Mn_2O_3 and HfO_2 phases. Consequently, we attempted to prepare samples in the form of thin films. In this case, we obtained a cubic Hf/MnO_2 phase. EDS analysis revealed that the oxygen concentration corresponds to an occupancy of 0.75. To verify whether the oxygen vacancies are a result of low oxygen concentration in the reservoir during thin-film growth, or if they are energetically driven by the specific Mn/Hf ratio, we again employed a universal potential to calculate the energies for various compositions and atomic distributions. Starting from a cubic HfO_2 structure, we replaced 50% of the Hf with Mn and considered various vacancy concentrations. Using the Supercell program, we generated all possible atomic configurations on appropriately sized supercells. With a finite (and in our case, small) number of atoms, the structure no longer maintains cubic or pseudo-cubic symmetry. To simulate real structures where cubic symmetry emerges statistically over a large number of atoms, it is appropriate to apply certain constraints. *Fig. 4* shows the formation energy of the Hf/MnO_2 structure as a function of oxygen vacancy concentration under different constraints during geometry optimization. It is evident that, regardless of the constraints applied, an oxygen occupancy around 0.75 is energetically favorable. Furthermore, it can be seen that the local ordering also has a significant impact on the total energy.

1. https://axelvandewalle.github.io/www-avdw//atat/
2. https://atat-sqs.streamlit.app/ .
3. https://orex.github.io/supercell/
4. https://github.com/bracerino/uMLIP-Interactive
5. https://xrdlicious-point-defects.streamlit.app/

L11

AZIMUTHAL INTEGRATION OF DIFFRACTION IMAGES: GEOMETRY, PHYSICAL CORRECTIONS, AND NXAZINT FORMAT

Z. Matěj[1], F. H. Gjørup[1,2], M. Yazdi-Rizi[1], P. Bell[1], W. De Nolf[3], M. R. V. Jørgensen[1,2]

[1]*MAX IV Laboratory, Lund University, Lund, Sweden*
[2]*Department of Chemistry & iNANO, Aarhus University, Aarhus, Denmark*
[3]*ESRF – The European Synchrotron, Grenoble, France*
zdenek.matej@maxiv.lu.se

Powder diffraction experiments at accelerator-based light sources predominantly employ transmission geometries with flat (2D) area detectors positioned downstream of the sample. This configuration enables efficient recording of complete Debye–Scherrer rings, taking advantage of the high diffracted intensities at low and intermediate scattering angles. When combined with large, fast detectors, it supports high-throughput as well as time-resolved powder diffraction measurements.

The use of 2D detectors preserves the full azimuthal distribution of the diffracted intensity, providing access to information on preferred orientation, grain statistics, and other deviations from ideal powder averaging. For routine structural analysis, detector images are typically reduced by azimuthal integration (AZINT) to yield one-dimensional diffraction profiles that follow the conventional powder diffraction formalism and serve as standard input for Rietveld refinement [1] and other quantitative methods. When deviations from ideal powder averaging are significant, azimuthally resolved integration or two-dimensional intensity representations can be employed, retaining directional information that enables a more comprehensive characterization of texture and related effects.

Azimuthal integration can be performed using several established software packages, including *fit2d* [2], *PyFAI* [3], *diffpy.srxplanar* [4], *Nika* [5], and the *AZINT module* [6]. These software tools or processing pipelines typically provide a wide range of configuration options. Core corrections commonly include solid-angle and polarization corrections, as well as intensity normalization. Additional, more specialized corrections may be applied when required, such as absorption or scattering effects arising from the sample, sample holder, or sample environment, as well as parallax corrections. More advanced functionality may include dedicated normalization and error-propagation models, along with treatments accounting for correlations between neighbouring points in powder diffraction patterns [4]. Complementing AZINT software, a variety of mature programs are available for subsequent data analysis steps; a non-exhaustive list of tools commonly used for Rietveld refinement includes *GSAS-II* [7], *TOPAS* [8], *Jana2020* [9], and *MStruct* [10].

Powder diffraction patterns and AZINT data have historically been stored in a wide range of predominantly text-based formats. However, modern time-resolved or spatially resolved experiments [11] at synchrotron light sources can practically produce hundreds of thousands to millions of diffraction patterns within minutes, making data handling and management increasingly challenging. To address this, most light sources worldwide have adopted the HDF5 data format as a versatile container for large volumes of hierarchically organized data and metadata. In addition, several diffraction instruments employ the NeXus [12] data standard built on top of HDF5, with the NXmx [13] application definition representing the current gold standard for macromolecular crystallography [14].

The first part of this contribution reviews the basic principles of azimuthal integration and the most commonly applied corrections used during data reduction. This discussion provides the foundation for the second part, which introduces the recently developed NXazint [15,16] format (Fig. 1). NXazint is designed to store 1D and 2D azimuthally integrated data together with essential metadata, enabling direct use in downstream analysis software. The format adheres to FAIR data principles, facilitates the handling of large datasets, and improves interoperability across beamlines and facilities. NXazint is already deployed at the MAX IV synchrotron, and broader adoption by other facilities and data analysis software is encouraged. The format is not intended to replace the pdCIF standard [17] but rather to serve as a complementary data container.

1 H. M. Rietveld, *J. Appl. Cryst.*, **2**, (1969), 65. doi:10.1107/S0021889869006558.

2. A. P. Hammersley, S. O. Svensson, M. Hanfland, A. N. Fitch, D. Hausermann, *High Press. Res.*, **14**, (1996), 235. doi:10.1080/08957959608201408.

3. J. Kieffer, J. P. Wright, *Powder Diffraction*, **28 s2**, (2013), 339. doi:10.1017/S0885715613000924.

4. X. Yang, P. Juhas, S. J. L. Billinge, *J. Appl. Cryst.*, **47**, (2014), 1273. doi:10.1107/S1600576714010516.

5. J. Ilavsky, *J. Appl. Cryst.*, **45**, (2012), 324. doi:10.1107/S0021889812004037.

6. *azint: python module for Azimuthal Integration*, https://maxiv-science.github.io/azint/, (April 29th, 2026).

7. B. H. Toby, R. B. Von Dreele, *J. Appl. Cryst.*, **46**, (2013), 544. doi:10.1107/S0021889813003531.

8. A. A. Coelho, *J. Appl. Cryst.*, **51**, (2018), 210. doi:10.1107/S1600576718000183.

9. V. Petříček, L. Palatinus, J. Plášil, M. Dušek, *Z. für Kristall. - Cryst. Mat.*, **238**, (2023), 271. doi:10.1515/zkri-2023-0005.

```
root:NXroot
  entry:NXentry
    @default = 'data'
    azint1d:NXsubentry
      @default = 'data'
      data:NXdata
        @axes = ['.', 'radial_axis']
        @interpretation = 'spectrum'
        @signal = 'I'
        I = float32(91x2000)
          @long_name = 'intensity'
          @units = 'arbitrary units'
        norm = float32(2000)
          @long_name = 'effective number of pixels contributing to the...'
          @units = 'arbitrary units'
        radial_axis = float64(2000)
          @long_name = '2theta'
          @units = 'degrees'
        radial_axis_edges = float64(2001)
          @long_name = '2theta bin edges'
          @units = 'degrees'
      definition = 'NXazint1d'
      instrument:NXinstrument -> /entry/instrument
        @default = 'name'
        monochromator:NXmonochromator
          @default = 'energy'
          energy = 34.99885420377407
            @units = 'keV'
          wavelength = 0.35424016820136567
            @units = 'angstrom'
        name = 'DanMAX'
        source:NXsource
          @default = 'name'
          name = 'MAX IV'
          probe = 'X-ray'
          type = 'Synchrotron X-ray Source'
      monitor:NXmonitor -> /entry/monitor
        data = float64(91)
```

Figure 1. Screenshot of a fragment of the NXazint formatted file, illustrating how 1D and 2D azimuthally integrated data and associated metadata are structured within the NeXus.

10. Z. Matěj, A. Kadlecová, M. Janeček, L. Matějová, M. Dopita, R. Kužel, *Powder Diffraction*, **29 s2**, (2014), 35. doi:10.1017/S0885715614000852.
11. E. K. Christensen, T. Moriishi, T. Komori, *Faraday Discussions*, **261**, (2025), 446. doi:10.1039/d5fd00009b
12. M. Konnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, S. I. Campbell, B. Clausen, S. Cottrell, J. U. Hoffmann, P. R. Jemian, D. Mannicke, R. Osborn, P. F. Peterson, T. Richter, J. Suzuki, B. Watts, E. Wintersberger, J. Wuttke, *J. Appl. Cryst.*, **48**, (2015), 301doi:10.1107/S1600576714027575
13. *NXmx: functional application definition for macromolecular crystallography*, https://manual.nexusformat.org/classes/applications/NXmx.html, (April 29th, 2026).
14. B. J. Bernstein, A. Förster, A. Bhowmick, A. S. Brewster, S. Brockhauser, L. Gelisio, D. R. Hall, F. Leonarski, V. Mariani, G. Santoni, C. Vonrhein, G. Winter, *IUCr-J*, **7**, (2020). 784. doi:10.1107/S2052252520008672
15. *NXazint1d: Application definition for data from two-dimensional area detectors that has been integrated azimuthally …*, https://manual.nexusformat.org/classes/applications/NXazint1d.html, (April 29th, 2026).
16. R. V. Jørgensen, F. H. Gjørup, M. Yazdi-Rizi, Z. Matej, W. De Nolf, P. Bell, J. *Synchrotron Rad.*, **33**, (2026), 896. doi:10.1107/S1600577526004224.
17. *POWDER CIF DICTIONARY*, https://www.iucr.org/resources/cif/dictionaries/cif_pd, (April 29th, 2026).