Rashchenko S.V. (2025) *J. Appl. Cryst.* **58**. https://doi.org/10.1107/S1600576724011956

crystchemlib — a Python library and GUI for analysis of crystal structure datasets

Sergey V. Rashchenko

Sobolev Institute of Geology and Mineralogy, Siberian Branch of Russian Academy of Sciences

rashchenko@igm.nsc.ru

Abstract

A problem of extracting basic (*e.g.*, bond lengths and angles) and advanced (*e.g.*, polyhedra volumes and effective coordination numbers) crystal chemical parameters from large datasets of CIFs in a quick and flexible way is addressed by a lightweight Python library with graphical user interface (GUI). A description of library functionality in GUI and scripting modes followed by examples based on open-access data demonstrates its advantages for crystallographers working with pressure-, temperature- and chemistry-induced structural variations, as well as with analysis of structural databases.

Keywords: *Python, database, polyhedron, bond length, pressure, temperature, CIF*

Introduction

The current progress in the technology of 4^{th} generation synchrotron sources and x-ray free-electron lasers together with enhancement of hybrid pixel detector capabilities inevitably lead to exponential growth of measured high-quality x-ray diffraction data (Kroon-Batenburg et al., 2024). Although much successful efforts has been made in the field of fast and efficient raw data processing to obtain structural information (*i.e.,* parameters of unit cell and atomic sites), the following step – crystallographic and crystal chemical analysis of the obtained structures compatible with large datasets – remains less addressed by developers of crystallographic software. A particular case to be dealt with in this work is large structural datasets, corresponding either to a single compound measured at varying parameter (pressure, temperature, time, external field, absorbed dose *etc.*) or to a compilation of structures with different composition but sharing a common structural feature.

Since standardized description of crystal structures in Crystallographic Information File (CIF) format is widespread and matured (Hall et al., 1991), the above task can be addressed by software with the following capabilities:

- reading ('parsing') of structural information from provided CIFs (including cases with many datablocks per file)

- calculation of user-defined parameters, characterizing a certain structural feature for each parsed datablock
- generation of ‘parameter vs. variable’ dependencies for selected parameters and such variables as pressure, temperature, time, external field, absorbed dose, chemical composition *etc.*

Although many software and libraries currently exist for CIF parsing and validation (see https://www.iucr.org/resources/cif/software), a crystallographer without programming skills is still forced to manually inspect tens of CIFs or CIF datablocks to extract *e.g.*, dependence of a certain bond length or unit cell parameter on pressure or temperature. The problem becomes even worse when the required numeric parameter is not present in CIF and should be calculated using a separate software. The latter is particularly relevant when extracting bond lenghts and angles not specified in original CIF, or when such complex quantities as coordination polyhedron volume or effective coordination number, not supported by CIF language, are of interest.

To address the aforementioned problem, a compact Python library *crystchemlib* capable of both CIF parsing and evaluating of basic structural features is proposed. The use of Python enables easy integration with a plenty of available data analysis and scientific plotting Python libraries, whereas a user-friendly browser graphical interface (GUI) provides access for crystallographers without programming skills. Below, the use of GUI will be first discussed for two examples from published datasets, then a brief description of *crystchemlib* core library will be given also followed by an example of use. The *crystchemlib* core library and GUI together with installation instructions are available at GitHub (https://github.com/SergeyRa/crystchemlib).

Graphical user interface (GUI)

The GUI of *crystchemlib* is based on *streamlit* library for browser applications and launches in the system default browser. In the main mode (‘crystchemlibGUI’) three interactive panels are available: ‘file import’, ‘polyhedra’, and ‘variables’ (Fig. 1).

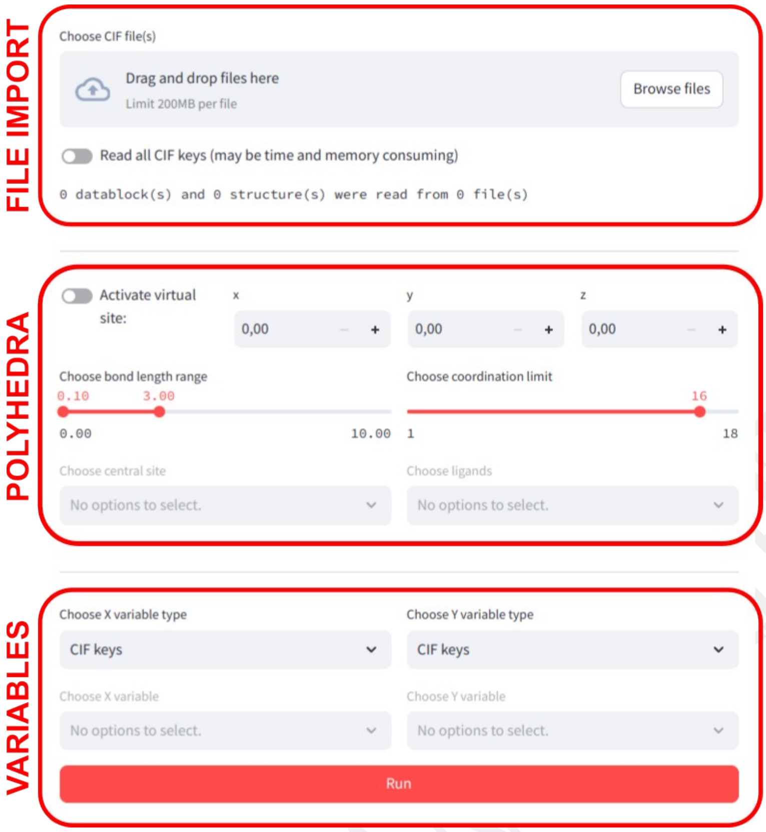

Figure 1. *Main panels of crystchemlib GUI*

'File import' panel enables selection of CIFs and supports CIFs with single and multiple datablocks per file, as well as their combination. By default only CIF keys necessary to build a structure model and those containing information about pressure and temperature of the experiment are parsed; *Read all CIF keys* switch forces program to parse all keys present in CIF.

'Polyhedra' panel provides settings for coordination polyhedra to be analyzed in the imported crystal structure data. To define a polyhedron, one needs to select central site and ligands, as well as specify distance range for bond search and maximum number of ligands (starting from the nearest one) to be considered. In cases when volume of an 'empty' polyhedron or structural void needs to be estimated, one can activate a 'virtual' site with user-defined coordinates and then use it as a central one.

The key panel is 'variables' which enables selection of structural parameter to be evaluated ('Y variable') and 'external' parameter (usually pressure, temperature, or chemical composition) to be used as 'X variable'. Four groups of variables are available for selection:

- *CIF keys:* enables selection of particular key (present in CIF) that directly specifies a desired X (*e.g.*, *_cell_measurement_pressure*) or Y (*e.g.*, *_cell_volume*) variable

- *CIF loops* (only for 'Y variable'): enables selection of key from CIF 'loops', which corresponds to a column of values (*e.g.*, a fractional coordinate). When such a key is selected, one can also select keys from the same loop to be used as labels in tables and plots (*e.g.*, site labels for bond lengths).
- *Formula content:* enables selection of element content extracted from *_chemical_formula_sum* CIF key (usually used as 'X variable' in analysis of 'chemical' deformations of crystal structure)
- *Other:* enables selection of the following coordination polyhedron parameters:
  - *Central site occupancy:* occupancy of the central site
  - *Coordination number:* number of ligands within specified distance range
  - *Effective coordination number:* ECoN, sum of 'bond weights' assuming the polyhedron to be homoligand (Hoppe, 1979; Nespolo, 2016)
  - *Mean distance:* mean L_i-C distance in the polyhedron where L is ligand and C – the central site
  - *Number of hidden ligands:* difference between total number of ligands found within the specified distance range and number of vertices in the convex hull formed by these ligands. Usually the presence of 'hidden' ligands indicates that the specified distance range extends to the second coordination sphere.
  - *Polyhedron angles:* L_i-C-L_j angles in the polyhedron where L is ligand and C – the central site
  - *Polyhedron bond weights:* weights of individual bonds according to 'charge distribution' formalism assuming the polyhedron to be homoligand (Hoppe, 1979; Nespolo, 2016)
  - *Polyhedron distances:* L_i-C distances in the polyhedron where L is ligand and C – the central site
  - *Polyhedron distances (corr.):* distances corrected for thermal motion using 'simple rigid bond' approach by (Downs et al., 1992)
  - *Polyhedron volume:* volume of the selected coordination polyhedron
  - *Polyhedron volume (corr.):* volume corrected for thermal motion using 'simple rigid bond' approach

The group 'Other' also contains two special variables '*P*, GPa' and '*T*, °C', which are just pressure and temperature converted from native CIF units (kPa and K) into GPa and °C. As soon the CIFs are downloaded and ‘X variable’ and ‘Y variable’ selected, the corresponding interactive plot and table can be produced by pressing ‘Run’ button (see examples below).

Example 1: Potassium coordination in maruyamaite under high pressure

Original publications on high-pressure or high-temperature crystallography usually concentrate only on certain aspects of a crystal structure, so that often an involved reader or reviewer would also like to check other aspects and hypotheses. The proposed GUI efficiently simplifies the latter task. In the example below open-access CIFs with maruyamaite (K-tourmaline) crystal structure compressed up to 20 GPa from electronic supplement to the paper of Likhacheva *et al.* (Likhacheva et al., 2019) are used to unveil the pressure-induced changes in the K^+ coordination not covered in the original work (Fig. 2).

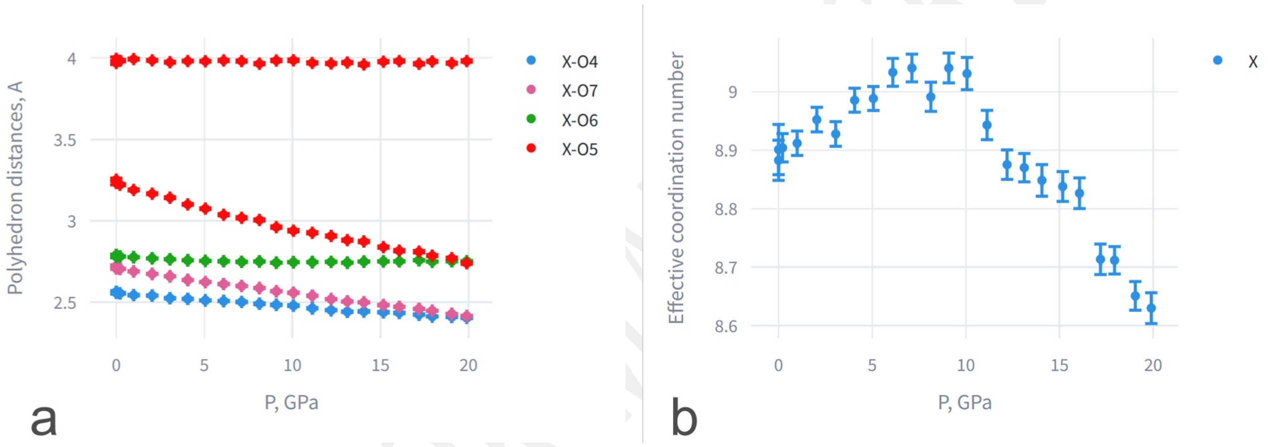

Figure 2. *Interactive plot output for distances (a) and ECoN (b) for the X-site coordination polyhedron in maruyamaite*

The distances and ECoN shown in Fig. 2 were plotted for the central site X (conventional label for the largest cation site in tourmaline structure) and O1-O8 sites selected as possible ligands; maximum bond length to be considered was set to 4.0 Å. All plotted points and their standard uncertainties are also printed below the plot as an interactive table, which can be easily exported in *.csv* format for further analysis and plotting.

As one can see in the Fig. 2a, the treatment of potassium (X) site coordination as a nine-fold in all explored pressure range, used in the original paper, is only partially correct. At lower pressures, the nine shortest bonds (3×X-O4, 3×X-O7, and 3×X-O6) indeed compose the first coordination sphere, which is also supported by the value of ECoN (Fig. 2b). However, with increasing pressure, the three O6 ligands move away from the first coordination sphere, while a single O5 site, situated in a special position, moves towards the first coordination sphere, so that at pressures above ~10 GPa, the coordination of X site should be described as 6+4, which also manifests in the pressure behavior of ECoN (Fig. 2b).

Example 2: 'Chemical' deformations in magnesite-otavite solid solution

(Bromiley et al., 2007) reported crystal structures of 27 intermediate compositions of magnesite-otavite ($MgCO_3$ – $CdCO_3$) solid solution (including ordered and disordered ones), currently available via American Mineralogist Crystal Structure Database (Downs & Hall-Wallace, 2003). The original paper, however, does not provide the dependence of cation octahedra volume *vs.* chemical composition, which can be easily obtained using *crystchemlib* GUI (Fig. 3).

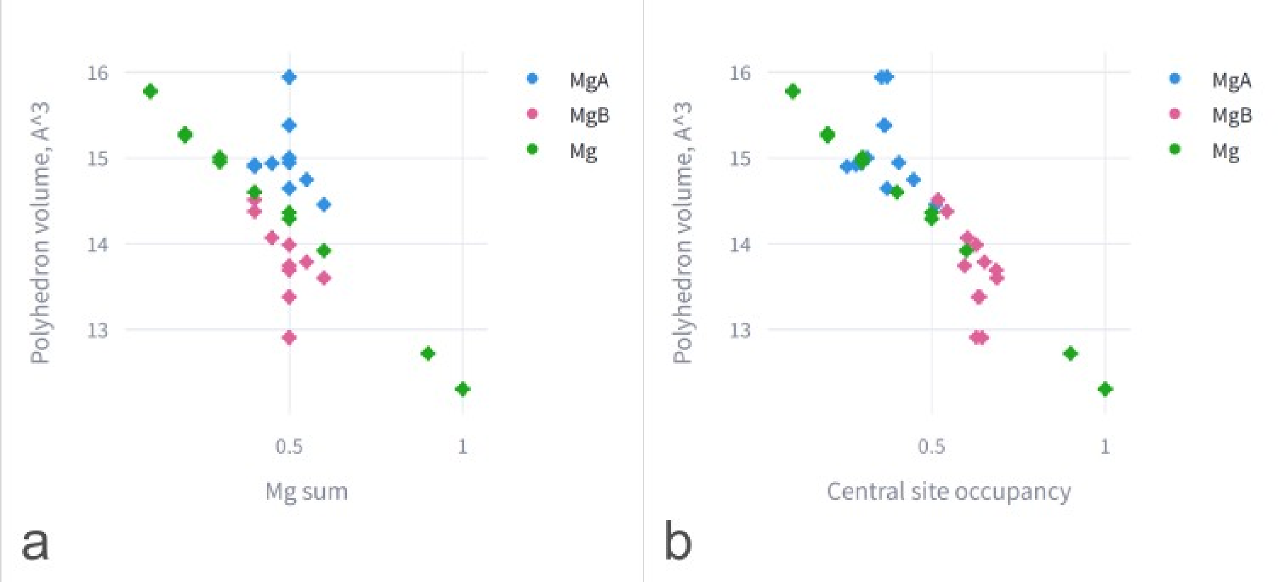

Figure 3. *Interactive plot output for volume of cation octahedra in ordered (MgA and MgB sites) and disordered (Mg site) members of magnesite-otavite solid solution, plotted versus total (a) and site-specific (b) Mg content*

In the Fig. 3a, the dependence of cation octahedra volume ('Mg' for disordered structures, 'MgA' and 'MgB' for ordered ones) is shown *versus* total magnesium content extracted from *_chemical_formula_sum* CIF key, so that a linear trend appears for disordered structures, characterized by a single cation site (Mg). In contrast, in ordered structures stable near Mg:Cd ~ 1:1, the same total Mg content may be distributed between two cation sites (MgA and MgB) in various ways, so that no evident correlation can be seen for MgA and MgB sites in Fig. 3a. To improve the presentation, one can switch the X variable to '*Other → Central site occupancy*', so that occupancy of Mg in individual polyhedra will be used as X variable instead of total Mg content, and a previously hidden trend will become evident (Fig. 3b).

Classes and functions of *crystchemlib* core library

The *crystchemlib* core library contained in *core.py* file is based on the following three classes (types of digital objects created during code execution):

- *Structure* – main class containing all information representing a given crystal structure (*i.e.* cell dimensions, symmetry operations and list of crystallographic sites)
- *Site* – class containing information about a crystallographic site (*i.e.* label, chemical symbol, fractional coordinates, occupancy and $U_{iso/eq}$)
- *Polyhedron* – miscellaneous class defining a coordination polyhedron

Each class is characterized by ‘attributes’ (values of its intrinsic parameters) and ‘methods’ (procedures that can be called to convert class attributes into derivative quantities or other output of interest) – see Table 1.

Table 1. *Classes of crystchemlib core library and their attributes and key methods*

Class	Attributes	Key methods
Structure	*cell* cell parameters [a, b, c, α, β, γ] *cell_esd* uncertainties of cell parameters [σ_a, σ_b, σ_c, σ_α, σ_β, σ_γ] *symops* list of symmetry operations in the form of 4×4 augmented matrices (Müller, 2013) *sites* list of Site objects	*cif* Outputs content of Structure in CIF format *formula* Returns formula unit *p1* Returns geometrically equivalent structure with *P*1 space group *poly* Creates Polyhedron object with given central site and ligands *sublatt* Returns substructure of specified sites *transform* Changes crystallographic basis
Site	*fract* fractional coordinates [x, y, z, 1]	

	fract_esd uncertainty of fractional coordinates [σ_x, σ_y, σ_z, 0] *label* label *symbol* chemical symbol *occ* occupancy *occ_esd* occupancy uncertainty *u* $U_{iso/eq}$ *u_esd* uncertainty of $U_{iso/eq}$	
Polyhedron	*central* central site of the polyhedron (Site object) *ligands* list of ligands (Site objects) *cell* cell parameters [a, b, c, α, β, γ] *cell_esd* uncertainties of cell parameters [σ_a, σ_b, σ_c, σ_α, σ_β, σ_γ]	*bondweights* Returns bond weights *econ* Returns effective coordination number *hidden* Returns number of hidden ligands *listangl* Returns angles in polyhedron *listdist* Returns distances in polyhedron *meandist* Returns mean distance in polyhedron

		polyvol Returns polyhedron volume

A particular attention was paid to estimation of uncertainties of such complex parameters as polyhedron volume, bond weights and ECoN, which are required for reliable analysis of their dependencies on pressure, temperature *etc.*, but often remain ignored in the dedicated software. For the uncertainty of polyhedron volume the approximation $\sigma_V = 3 \cdot \sigma_r$ is used, where σ_V is *relative uncertainty* of the polyhedron volume, and σ_r is a *mean relative uncertainty* of distances between the polyhedron vertices and its geometrical center. For the uncertainty of bond weight the approximation $\sigma_w = 6 \cdot \sigma_d \cdot (1 - \ln w)$ is used, in which σ_w is a *relative uncertainty* of the bond weight, σ_d is a *relative uncertainty* of the corresponding bond length, and w is the bond weight itself. The squared absolute uncertainty of ECoN is just the quadratic sum of absolute uncertainties of the bond weights, accounted in the ECoN.

The *crystchemlib* core library also contains a number of 'functions' – minor crystallographic routines extensively used throughout the code. The most important ones are:

- *angle* (returns an angle between two vectors defined by fractional coordinates in a given unit cell)
- *dhkl* (returns an interplanar distance for given *hkl* and unit cell)
- *equivhkl* (returns a list of symmetrically equivalent *hkl* indices for given unit cell and symmetry operations)
- *length* (returns a distance between two points defined by fractional coordinates in a given unit cell)
- *newbasis* (returns [a, b, c, α, β, γ] transformed by a given transformation matrix)
- *orthonorm* (transforms fractional coordinates in a given unit cell into orthonormal ones following McKie (McKie & McKie, 1986))
- *parsecif* (extracts CIF content into a Python 'dict' object)
- *readstruct* (returns a Structure object from a CIF-based Python 'dict' object)
- *vol* (calculates a unit cell volume)

An advanced user can find more documentation on the described (and many other) features of *crystchemlib* core library inside the code (*core.py* file) in the format of so called *docstrings*.

Example 3: Boron – oxygen bond length statistics from Crystallography Open Database

Often in structural studies one is interested in 'typical' interatomic distances for a given pair of atoms and its statistical variation (preferably for a given class of compounds). Although the corresponding data occasionally appear in relevant reviews, the increasing number of crystal structures deposited in databases makes crystallographers wish for an instrument that enables extraction of basic structural features from user-defined structural sets in an easy, quick and flexible way. In this example, the capabilities of *crystchemlib* core library are used to extract B–O bond length statistics for single- and double-cation borates deposited in Crystallography Open Database (Gražulis et al., 2009). An initial CIF dataset, containing 1245 files, was downloaded using web search interface and following constraints:

- Element 1: B
- Element 2: O
- Not these elements: C, H
- Number of distinct elements: from 3 to 4

A short Python script exploiting the capabilities of *crystchemlib* core library was prepared to extract from the dataset all B–O bond lengths falling in the range of 1-2 Å and plot them as a histogram (Fig. 4). After execution, the resulting histogram was plotted (Fig. 5), demonstrating bimodal distribution with modes near 1.37 Å and 1.47 Å, evidently corresponding to sp^2- and sp^3-bonded boron, respectively.

```python
import core
import matplotlib.pyplot as plt
import numpy as np
import os


dir = './test_datasets/6_COD_BO_notCH_3-4/'
cifs = [i for i in os.listdir(dir) if i.endswith('.cif')]

structures = []
for c in cifs:
    with open(dir+c) as f:
        parsed = core.parsecif(f)
        for j in parsed['data']:
            struct = core.readstruct(j)
            if struct is not None:
                structures.append(struct)

polyhedra = []
for s in structures:
    centr = s.filter('symbol', ['B', 'B3+'])
    ligands = s.filter('symbol', ['O', 'O2-'])
    for j in centr:
        polyhedra.append(s.poly(j, ligands, 2, dmin=1))

distances = []
for p in polyhedra:
    distances += p.listdist()['value']

fig, axes = plt.subplots(dpi=200)
axes.hist(np.array(distances), bins='auto')
axes.set_xlabel('Bond length, $\AA$')
axes.set_ylabel('Number of bonds')
```

Figure 4. *Python script written to extract B–O bond length statistics for single- and double-cation borates from the Crystallography Open Database*

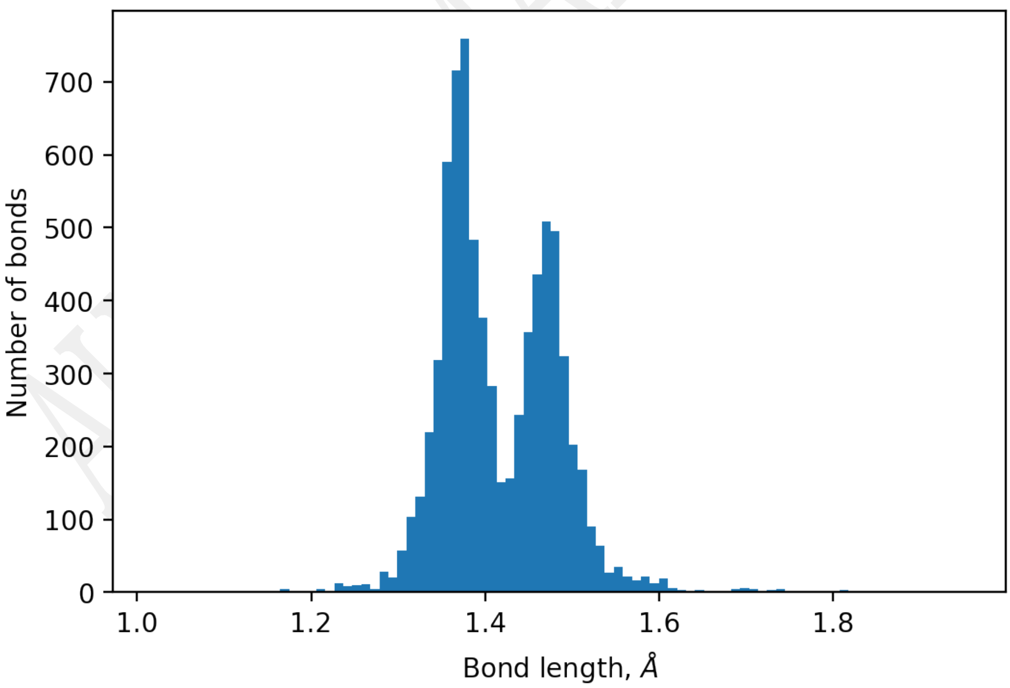

Figure 5. *B–O bond length histogram for single- and double-cation borates from the Crystallography Open Database*

Conclusions

The proposed Python library and GUI provides an open-source solution for crystal chemical analysis of both user data and published datasets. A user-friendly GUI ensures availability of the basic library features for crystallographers without programming skills, whereas more advanced use is possible for those familiar with Python.

Acknowledgments

The work was supported by the Russian Science Foundation (grant #23-77-10047). Dr. Liudmila Gorelova, Dr. Oleg Vereshchagin (Saint-Petersburg State University), Mark Ignatov, and Alexandr Romanenko (Sobolev Institute of Geology and Mineralogy) are acknowledged for providing suggestions, test datasets, and testing of the software.

References

Bromiley, F. A., Ballaran, T. B., Langenhorst, F. & Seifert, F. (2007). *Am. Mineral.* **92**, 829–836.

Downs, R. T., Gibbs, G. V. & Bartelmehs, K. L. (1992). *Am. Mineral.* **77**, 751–757.

Downs, R. T. & Hall-Wallace, M. (2003). *Am. Mineral.* **88**, 247–250.

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.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hoppe, R. (1979). *Z. Kristallogr. – Cryst. Mater.* **150**, 23–52.

Kroon-Batenburg, L. M. J., Lightfoot, M. P., Johnson, N. T. & Helliwell, J. R. (2024). *Struct. Dyn.* **11**, 011301.

Likhacheva, A. Yu., Rashchenko, S. V., Musiyachenko, K. A., Korsakov, A. V., Collings, I. E. & Hanfland, M. (2019). *Miner. Petrol.* **113**, 613–623.

McKie, D. & McKie, C. (1986). *Essentials of crystallography.* Oxford: Blackwell Scientific Publications.

Müller, U. (2013). *Symmetry relationships between crystal structures: applications of crystallographic group theory in crystal chemistry.* Oxford University Press.

Nespolo, M. (2016). *Acta Cryst.* B**72**, 51–66.