The Journal of Open Source Software

crystract: A Crystallography Package in R for .cif Data Processing

Don Ngo [1¶], **Julia M. Hübner** [2], **Marc Spitzner**[3], **Shaunna M. Morrison** [4], **and Anirudh Prabhu** [1¶]

1 Earth and Planets Laboratory, Carnegie Institution for Science, Washington, DC, USA **2** Technische Universität Dresden, Dresden, Germany **3** Independent Scholar Dresden, Germany **4** Department of Earth and Planetary Sciences, Rutgers University, Piscataway, NJ, USA ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Bonan Zhu

Reviewers:

- @bobleesj
- @singularitti
- @hisacro

Submitted: 16 September 2025
Published: unpublished

Summary

The Crystallographic Information File (CIF) is the standard format for disseminating crystal structure data, yet parsing and analyzing these files for large-scale computational and statistical analysis is a significant bottleneck for research in the chemical and material sciences (Hall et al., 1991). `crystract` is an R package designed to provide an efficient, open-source solution for the batch processing and statistical analysis of CIF files. The package streamlines the extraction of metadata, unit cell parameters, atomic coordinates, and symmetry operations for a singular or hundreds of CIF files at the same time. From the information stored in the CIF file, our package can determine the first-neighbor bonding environment around each symmetrically independent atom in a unit cell, with all interatomic distances and bond angles, while propagating experimental uncertainties. Furthermore, a comprehensive workflow for efficient extraction and processing of these data is provided, centered around the calculation of average interatomic distances, as identified as a key parameter for a streamlined comparison of structural features of different compounds. One of the most important features within this workflow is the package's ability to process positional or occupational disorder. To handle positional disorder, a filtering function to eliminate non-physical distances is provided. Occupational disorder is taken into account via the possibility of calculating an occupancy-weighted average interatomic distance. Additionally, filtering functions to calculate average distances, only including user-specified elements or atomic positions, are available. This paper outlines the architecture and core functionalities of `crystract`, demonstrating its utility with a practical example.

Statement of need

Despite the standardization provided by the CIF format, significant practical barriers remain for researchers aiming to perform high-throughput computational analysis, particularly in batch. The first barrier is technical: CIF files, while standardized, often exhibit syntactic variations or errors depending on their originating software or laboratory, which can cause simplistic parsers to fail. The second barrier is conceptual: a CIF file does not typically contain an explicit list of all atoms in the unit cell. Instead, it reports the unique atoms in the asymmetric unit and a set of symmetry operations. A complete structural analysis, therefore, requires the correct application of these symmetry operations to generate the full unit cell and its atomic contents—a non-trivial, error-prone computational task that must be handled by specialized software. This complexity often forces researchers into a fragmented and inefficient workflow, piecing together disparate tools for data validation, structure generation, geometric analysis, and final statistical modeling.

A number of excellent software tools have been developed to address some of these challenges, yet a comprehensive review reveals a specific and critical gap. The most mature ecosystem for computational materials science currently resides in Python, where `pymatgen` stands as a powerful and widely adopted standard (Ong et al., 2013). This extensive library offers robust CIF parsing and a host of advanced analysis functions, including the calculation of interatomic distances and the prediction of bonding environments using sophisticated, data-mining approaches like the CrystalNN algorithm (Pan et al., 2021). While `pymatgen` offers robust CIF parsing and advanced analysis, it has limitations for a complete, statistically rigorous workflow: it does not natively compute bond angles, nor does it programmatically propagate the experimental uncertainties reported in CIFs for its derived geometric quantities. Furthermore, batch processing requires the user to write custom scripts to loop over files. Filtering presents another challenge: although `pymatgen` includes a function to filter for minimum and maximum distances, they have to be specified one element at a time, one file at a time, by the user. This makes it cumbersome to use `pymatgen` for large scale datasets and studies.

Other specialized Python tools like `cifkit`—a Python based command-line tool (Lee & Oliynyk, 2024)—and its companion software, `cif-bond-analyzer` (CBA)—designed for the singular task of exporting bond lists (Tyvanchuk et al., 2024)—provide fast, lightweight, and native batch processing for geometric analysis. However, their scope is intentionally limited to interatomic distances and bond analysis; they also do not provide native functions for calculating bond angles or propagating experimental uncertainties.

Other specialized tools, such as the CCDC's Mercury (Macrae et al., 2006) or the IUCr's enCIFer (Allen et al., 2004), are indispensable for interactive 3D visualization and formal CIF syntax validation, respectively. However, their primary design as graphical user interface (GUI) applications makes them ill-suited for the automated, scriptable, and reproducible workflows required for modern data science without the possibility of handling a high throughput.

Within the R ecosystem, the landscape is sparse. The `cry` package provides basic statistics for crystallography computation and falls short of large-scale research and ML based applications (Foadi et al., 2025). Its design, centered on a custom S3 object system, is not optimized for the high-throughput, data-frame-centric workflows required for large-scale statistical analysis. `cry` is limited to the analysis of crystallographic parameters and diffraction data from individual files. It is not equipped for the geometric analysis of atomic structures, as it is unable to apply symmetry operations to generate a full unit cell from asymmetric coordinates; therefore, unable to calculate the resulting interatomic distances and angles, or handle the structural disorder commonly found in real materials.

To our knowledge, no package or software—whether in Python, as GUIs, or within R—provides a single, integrated solution capable of combining the automated batch processing of large collections of CIF files and the systematic propagation of experimental uncertainties, while providing additional features indispensable for the structural analysis of large datasets. Such a research software landscape forces researchers into creating fragmented and inefficient workflows, piecing together disparate tools for structure generation, geometric analysis, and finally statistical modeling.

Overview of functionalities provided by existing packages in Python and R.

Task	CIFkit	CBA (CIF Bond analyzer)	pymatgen	cry	crystract
Read/parse singular CIF file	Yes	No, uses cifcit for this	Partially, via importers like CifParser	Yes	Yes

Task	CIFkit	CBA (CIF Bond analyzer)	pymatgen	cry	crystract
Batch / high-throughput processing of many CIFs	Yes	Yes	Yes, but certain tasks require manual scripting	No	Yes
Supercell / unit cell generation / lattice operations	Yes, can generate unit cell and supercell via +/- 1 shifts	Yes, when computing minimum bond lengths for site	Yes, structure operations, transformation, supercell, etc.	No	Yes, can generate unit cell and supercell via +/- 1 shifts
Coordination number determination	Yes	Yes	Yes	No	Yes
Calculation of interatomic distances	Yes	Yes	Yes	No	Yes
Calculation of bond angles	No	No	No	No	Yes
Error propagation	No	No	No	No	Yes
Handling of occupational disorder	Yes	No, uses cifkit	Partial, can read occupancies, but does not use them	No	Yes
Handling of structural disorder	No	No	No	No	Yes, via filtering function
Filtering for specific atoms or crystallographic sites	Yes	Yes	No	No	Yes
Calculation of weighted average accounting for disorder.	No	No	No	No	Yes
Output to multiple formats	No outputs of extracted data, only figures such as histograms or visualizations	No outputs of extracted data, only figures such as histograms or visualizations	No	No	Yes

Crystract

To address these needs, we have developed "crystract", an open-source R package designed to provide a seamless, robust, and statistically-minded workflow for crystallographic analysis. crystract provides an end-to-end toolkit that operates entirely within the R environment. Its primary contributions are fourfold.

First, it provides a robust and efficient engine for parsing and processing large batches of CIF files. It is designed from the ground up around R's data-centric paradigm, directly presenting all extracted and calculated data in tidy data frames (Wickham, 2014) ready for immediate manipulation and analysis with the wider R ecosystem.

Second, it offers comprehensive geometric analysis. This output includes not only the CIF file's core metadata but also a rich set of derived attributes essential for crystallographic research: a complete list of atomic coordinates after the application of symmetry operations to generate the full cell, all interatomic distances based on predicted bonded pairs for direct neighbors within the coordination sphere of an individual atom using the Crystal NN algorithm (Pan et al., 2021), and bond angles—a feature not natively available in any other command-line batch-processing tools.

Third, and most uniquely, crystract introduces a capability largely absent in other programmatic tools: the rigorous propagation of experimental uncertainties from the CIF through all derived geometric quantities. This feature, grounded in standard error propagation theory (K\upsilon, 1966) facilitates a more sophisticated and honest statistical treatment of structural data, allowing researchers to quantify the confidence in their calculated results.

Fourth, as the most valuable feature for the user community, it provides a suite of integrated filtering functions that are essential for high-throughput analysis and handling complex structures, thus exceeding the capabilities offered by currently existing software tools. Our workflow is centered around the calculation of average interatomic distances, as a key parameter in the structural comparison of different compounds. The average atomic distance can be calculated for all symmetrically independent atoms or a subset of these by prior application of the filtering function based on the user-specified element or atomic site. Furthermore, occupational or positional disorder can be handled via a filtering function to exclude non-physical distances in the process of calculating the average distance. This filtering is based on the automatic recognition of the elements in a structure and the calculation of an expected interatomic distance from their covalent radii (Emsley, 1998; Pyykkö & Atsumi, 2009) or a user-specified list of atomic radii. Partial site occupation occurs in many real compounds and is indispensable for deriving structure-property trends or predicting new functional materials (Jakob et al., 2025). If one simply calculates interatomic distances from the coordinates given in the CIF file without accounting for partial occupation, one obtains non-physical distances between atoms that cannot coexist in the same local configuration. Such artifacts would distort any statistical measure by including interactions that never occur in reality. While taking these features native to real crystal structures into account, a weighted average distance can be calculated, either for all individual atoms or a user-defined subset by the application of the available filtering functions for user-specified atoms or crystallographic sites.

Implications

The availability of packages like "crystract", "pymatgen", "cifkit", and others bring the field of crystallography, mineralogy, and materials science a step closer to realizing the transformative potential of data-driven science. crystract is one part of a larger effort made to develop AI methods to perform comparative studies across hundreds or even thousands of structures thus providing researchers with the foundations to derive overarching structure-property relationships of minerals and materials to ultimately employ known compounds for new applications or predict new materials.

`crystract` creates a launchpad for integrating crystallographic data into machine learning pipelines for a variety of application areas. Our package's capability to handle positional or occupational disorder, propagate uncertainties, and generate comprehensive data outputs provide an excellent feature set for predictive modeling efforts.

Finally, as an open source package, "`crystract`" will be a community-driven, transparent resource that invites extensions and improvements from other researchers who want to use crystallographic data in their own scientific explorations.

Acknowledgements

The authors would like to thank Michael Baitinger, Robert T. Downs, Jolyon Ralph, and Xiaogang Ma for their discussions on crystallography, and cyberinfrastructure development. D.N. has been supported by the Earth and Planetary Science Interdisciplinary Internship at Carnegie Science (a National Science Foundation REU). Many thanks to Dionysis Foustoukos, Johanna Teske, Carnegie Science, and the National Science Foundation for the internship oppurtunity. Additionally, funding and support for this project was provided by the Carnegie Science and a private foundation.

Author Contributions using the CRediT (Contribution Roles Taxonomy)

Conceptualization – AP, JMH

Data Curation – DN, JMH, AP

Formal Analysis – DN, JMH, AP

Funding Acquisition – AP

Investigation – DN, JMH, AP

Methodology – DN, JMH, AP

Project Administrator – AP, JMH

Resources – SMM, JMH, AP, MS

Software – DN, JMH, AP

Supervision – AP, JMH

Validation – DN, JMH, SMM, AP, MS

Writing (Original Draft Preparation) – DN, JMH, AP

Writing (Review & Editing) – DN, JMH, AP, SMM, MS

References

Allen, F. H., Johnson, O., Shields, G. P., Smith, B. R., & Towler, M. (2004). CIF applications. XV. enCIFer: A program for viewing, editing and visualizing CIFs. *Applied Crystallography*, *37*(2), 335–338. https://doi.org/10.1107/S0021889804003528

Emsley, J. (1998). *The elements* (3rd ed.). Oxford University Press.

Foadi, J., Waterman, D., Giordano, R., & Nidamarthi, K. (2025). *Cry: Statistics for structural crystallography*. https://jfoadi.github.io/cry/

Hall, S. R., Allen, F. H., & Brown, I. D. (1991). The crystallographic information file (CIF): A new standard archive file for crystallography. *Foundations of Crystallography*, *47*(6), 655–685.

Jakob, K., Walsh, A., Reuter, K., & Margraf, J. T. (2025). Learning Crystallographic Disorder: Bridging Prediction and Experiment in Materials Discovery. *ChemRxiv*. https://doi.org/10.26434/chemrxiv-2025-f52qs

K\upsilon, H. (1966). Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards: Engineering and Instrumentation. Section C.*, *70*(4), 263.

Lee, S., & Oliynyk, A. O. (2024). Cifkit: A python package for coordination geometry and atomic site analysis. *Journal of Open Source Software*, *9*(103), 7205. https://doi.org/10.21105/joss.07205

Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M., & Streek, J. (2006). Mercury: Visualization and analysis of crystal structures. *Applied Crystallography*, *39*(3), 453–457. https://doi.org/10.1107/S002188980600731X

Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., & Ceder, G. (2013). Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, *68*, 314–319. https://doi.org/https://doi.org/10.1016/j.commatsci.2012.10.028

Pan, H., Ganose, A. M., Horton, M., Aykol, M., Persson, K. A., Zimmermann, N. E., & Jain, A. (2021). Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic Chemistry*, *60*(3), 1590–1603. https://doi.org/10.1021/acs.inorgchem.0c02996

Pyykkö, P., & Atsumi, M. (2009). Molecular single-bond covalent radii for elements 1–118. *Chemistry–A European Journal*, *15*(1), 186–197. https://doi.org/10.1002/chem.200800987

Tyvanchuk, Y., Babizhetskyy, V., Baran, S., Szytuła, A., Smetana, V., Lee, S., Oliynyk, A. O., & Mudring, A.-V. (2024). The crystal and electronic structure of $RE_{23}Co_{6.7}In_{20.3}$ (RE = Gd–Tm, Lu): A new structure type based on intergrowth of AlB_2- and CsCl-type related slabs. *Journal of Alloys and Compounds*, *976*, 173241. https://doi.org/10.1016/j.jallcom.2023.173241

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*, 1–23. https://doi.org/10.18637/jss.v059.i10