




# 1 crystract: A Crystallography Package in R for .cif 2 Data Processing

3 Don Ngo <sup>1¶</sup>, Julia M. Hübner <sup>2</sup>, Marc Spitzner<sup>3</sup>, Shaunna M.  
4 Morrison <sup>4</sup>, and Anirudh Prabhu <sup>1¶</sup>

5 <sup>1</sup> Earth and Planets Laboratory, Carnegie Institution for Science, Washington, DC, USA <sup>2</sup> Technische  
6 Universität Dresden, Dresden, Germany <sup>3</sup> Elbe Flugzeugwerke GmbH, Dresden, Germany <sup>4</sup> Department  
7 of Earth and Planetary Sciences, Rutgers University, Piscataway, NJ, USA ¶ Corresponding author

DOI: [10.xxxxxx/draft](https://doi.org/10.xxxxxx/draft)

## Software

- [Review](#) 
- [Repository](#) 
- [Archive](#) 

Editor: 

Submitted: 16 September 2025

Published: unpublished

## License

Authors of papers retain copyright  
and release the work under a  
Creative Commons Attribution 4.0  
International License ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/))

## 8 Summary

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

## 29 Statement of need

30 Despite the standardization provided by the CIF format, significant practical barriers remain for  
31 researchers aiming to perform high-throughput computational analysis, particularly in batch.  
32 The first barrier is technical: CIF files, while standardized, often exhibit syntactic variations or  
33 errors depending on their originating software or laboratory, which can cause simplistic parsers  
34 to fail. The second barrier is conceptual: a CIF file does not typically contain an explicit  
35 list of all atoms in the unit cell. Instead, it reports the unique atoms in the asymmetric unit  
36 and a set of symmetry operations. A complete structural analysis, therefore, requires the  
37 correct application of these symmetry operations to generate the full unit cell and its atomic  
38 contents—a non-trivial, error-prone computational task that must be handled by specialized  
39 software. This complexity often forces researchers into a fragmented and inefficient workflow,  
40 piecing together disparate tools for data validation, structure generation, geometric analysis,  
41 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., 2020). 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—and its companion software, cif-bond-analyzer (CBA)—designed for the singular task of exporting bond lists—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 (Groom et al., 2016) or the IUCr's enCIFer, 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 (Roveda, 2017). 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	crystrack
Read/parse singular CIF file	Yes	No, uses cifkit for this	Partially, via importers like CifParser	Yes	Yes

Task	CIFkit	CBA (CIF Bond analyzer)	pymatgen	cry	crysttract
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

## 84 Crystract

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

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

93 Second, it offers comprehensive geometric analysis. This output includes not only the CIF file’s  
94 core metadata but also a rich set of derived attributes essential for crystallographic research: a  
95 complete list of atomic coordinates after the application of symmetry operations to generate  
96 the full cell, all interatomic distances based on predicted bonded pairs for direct neighbors  
97 within the coordination sphere of an individual atom using the Crystal NN algorithm (Pan  
98 et al., 2020), and bond angles—a feature not natively available in any other command-line  
99 batch-processing tools.

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

105 Fourth, as the most valuable feature for the user community, it provides a suite of integrated  
106 filtering functions that are essential for high-throughput analysis and handling complex struc-  
107 tures, thus exceeding the capabilities offered by currently existing software tools. Our workflow  
108 is centered around the calculation of average interatomic distances, as a key parameter in the  
109 structural comparison of different compounds. The average atomic distance can be calculated  
110 for all symmetrically independent atoms or a subset of these by prior application of the filtering  
111 function based on the user-specified element or atomic site. Furthermore, occupational or  
112 positional disorder can be handled via a filtering function to exclude non-physical distances  
113 in the process of calculating the average distance. This filtering is based on the automatic  
114 recognition of the elements in a structure and the calculation of an expected interatomic  
115 distance from their covalent radii (Emsley, 1998; Pyykkö & Atsumi, 2009) or a user-specified  
116 list of atomic radii. Partial site occupation occurs in many real compounds and is indispensable  
117 for deriving structure-property trends or predicting new functional materials (Jakob et al.,  
118 2025). If one simply calculates interatomic distances from the coordinates given in the CIF  
119 file without accounting for partial occupation, one obtains non-physical distances between  
120 atoms that cannot coexist in the same local configuration. Such artifacts would distort any  
121 statistical measure by including interactions that never occur in reality. While taking these  
122 features native to real crystal structures into account, a weighted average distance can be  
123 calculated, either for all individual atoms or a user-defined subset by the application of the  
124 available filtering functions for user-specified atoms or crystallographic sites.

## 125 Implications

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

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

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

## 140 Acknowledgements

141 The authors would like to thank Michael Baitinger, Robert T. Downs, Jolyon Ralph, and  
142 Xiaogang Ma for their discussions on crystallography, and cyberinfrastructure development.  
143 D.N. has been supported by the Earth and Planetary Science Interdisciplinary Internship at  
144 Carnegie Science (a National Science Foundation REU). Additionally, funding and support for  
145 this project was provided by the Carnegie Institution for Science and a private foundation.

## 146 Author Contributions using the CRediT (Contribution Roles 147 Taxonomy)

148 Conceptualization – AP, JMH

149 Data Curation – DN, JMH, AP

150 Formal Analysis – DN, JMH, AP

151 Funding Acquisition – AP

152 Investigation – DN, JMH, AP

153 Methodology – DN, JMH, AP

154 Project Administrator – AP, JMH

155 Resources – SMM, JMH, AP, MS

156 Software – DN, JMH, AP

157 Supervision – AP, JMH

158 Validation – DN, JMH, SMM, AP, MS

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

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

## 161 References

162 Emsley, J. (1998). *The elements*. Oxford University Press.

163 Groom, C. R., Bruno, I. J., Lightfoot, M. P., & Ward, S. C. (2016). The Cambridge Structural  
164 Database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and*  
165 *Materials*, 72(2), 171–179. <https://doi.org/10.1107/s2052520616003954>

166 Hall, S. R., Allen, F. H., & Brown, I. D. (1991). The crystallographic information file (CIF): A  
167 new standard archive file for crystallography. *Acta Crystallographica Section A: Foundations*  
168 *of Crystallography*, 47(6), 655–685. <https://doi.org/10.1107/s010876739100531x>

- 169 Jakob, K., Walsh, A., Reuter, K., & Margraf, J. T. (2025). Learning Crystallographic  
170 Disorder: Bridging Prediction and Experiment in Materials Discovery. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv-2025-f52qs>  
171
- 172 Ku, H. H. (1966). Notes on the Use of Propagation of Error Formulas. *Journal of Research of*  
173 *the National Bureau of Standards, Section C: Engineering and Instrumentation*, 70C(4),  
174 263–273.
- 175 Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D.,  
176 & Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source  
177 Python library for materials analysis. *Computational Materials Science*, 68, 314–319.  
178 <https://doi.org/10.1016/j.commatsci.2012.10.028>
- 179 Pan, F., Mi, W., Gari, R. S. K., & Horton, M. (2020). Crystal structure generative modeling  
180 with deep learning. *Journal of Materials Chemistry A*, 8(11), 5576–5585. <https://doi.org/10.1039/C9TA01378D>  
181
- 182 Pyykkö, P., & Atsumi, M. (2009). Molecular Single-Bond Covalent Radii for Elements  
183 1–118. *Chemistry – A European Journal*, 15(1), 186–197. <https://doi.org/10.1002/chem.200800987>  
184
- 185 Roveda, J. (2017). *cry: Crystallography in R*. R package version 0.1.2.
- 186 Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), 1–23. <https://doi.org/10.18637/jss.v059.i10>  
187

DRAFT