Interatomic Distance List Database and Deep Learning for Ab Initio Structure Solution From PDF Data

Student: Sean Wu[1] Mentor: Simon J. L. Billinge[2]

[1]Pepperdine University, [2]Columbia University Department of Material Science and Engineering

Background

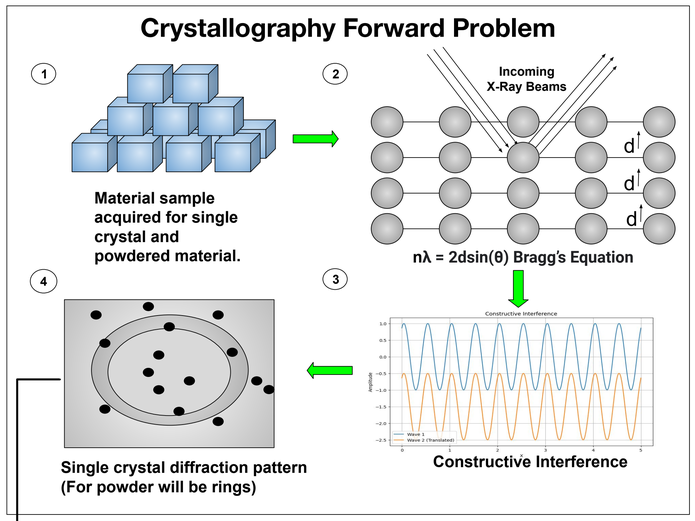


Figure: Illustration of the forward and inverse problems of crystallography. The forward process is a relatively trivial process where scientists pass x-ray beams into the material and record the scattering pattern. The inverse problem is not yet solved for all cases.

The inverse problem in crystallography [2] is one that gives incredible insight to a materials interatomic structure and makeup. In practice, solving the structure solution problem allows scientists from a myriad of disciplines, ranging from drug synthesis and even efficient battery design. As shown in the figure above, the forward process is relatively trivial, consisting of passing x-ray signals through a material to record the constructive interference on a scattering pattern. While the inverse problem has been solved for simple crystal structures, structure solutions from powder diffraction is much harder, and more advanced methods may be needed.

Problem Definition

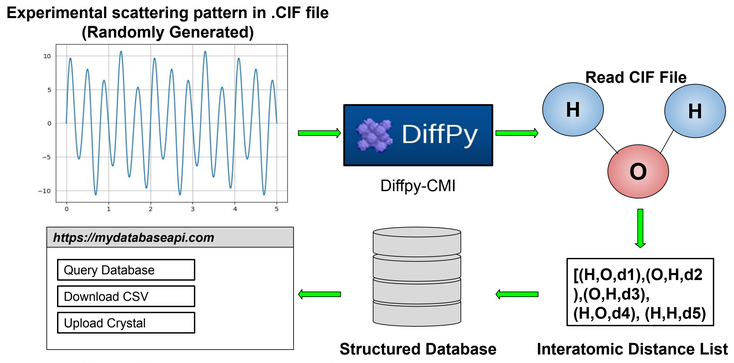
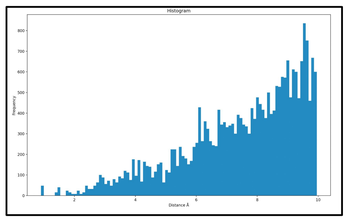


Figure: End to end representation of adding a new crystal to our existing database where we first pass the cif file into diffpy-cmi bond calculator, and we pass it through our script to append to the rest-api.

In more complex nano-materials, the one dimensional diffraction data is not enough to reconstruct the 3D electron density map. Therefore the pair distribution function is needed (PDF), which is the weighted histogram of interatomic distances. To compute the distance lists of an atom, a double sum is required, which can be translated into a nested for loop in coding with a $O(n^2)$ time complexity. When we compute these pair distribution functions for a high enough r(Å), this exponential growth can take a long time for researchers to devise data analysis. By pre-computing a distance list in a database, we allow for more intricate problems to be solved, because researchers can compute the pair distribution function in linear or even constant time. To do this we use the BondCalculator class from diffpy-cmi [3].

Results

In this study, we devised a python framework that not only builds, but maintains a distance list database, where researchers can add, create, and query new materials to the data. Although preliminary, we obtained 785 experimental .CIF files and wrote an extensive .json database that consists of not only the distance list but also other metadata of the material including the space group, and lots of information regarding the unit cell. As shown in the figure below, this is the result of the weighted histogram of the distance lists, that result in a radial distribution function (RDF).



+

$$G(r) = 4\pi r \rho_0 [g(r) - 1]$$

Figure: Visualization of the radial distribution function that is a histogram form of the distance lists of some material. As seen in the equation, we can obtain the pair distribution function parameterized by G(r) from R(r) [1]

Figure: Depiction of one material in our database with an abbreviated distance list. The three options are add, create, and query.

Discussion

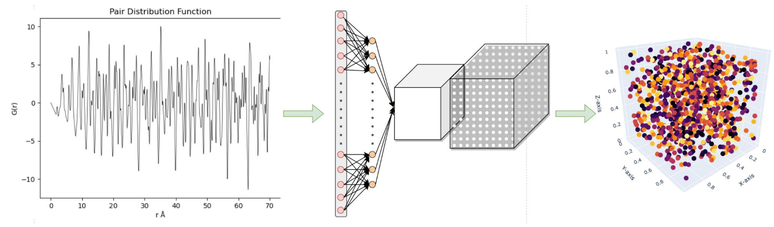


Figure: Figure of a sample use case for our project. Researchers can carefully curate a dataset of pair distribution functions from our distance lists, and generate a many PDFs to train a deep neural network to solve the inverse problem.

In this research, we created set the groundwork to solve two very fundamental and practical problems in crystallography. The first being the long computational time to compute a pair distribution function. With our precomputed distance list, it allows for researchers to use the distance lists for analyzing pair distributions functions are different experimental conditions such as heat and instrument parameters by convolving the delta functions with a gaussian distribution. Our group has many ambitions for future studies. Firstly, we hope to run our database maintainer on a large database of CIF files obtained from publicly available databases such as ICSD to fully publish the distance list of databases. We believe that a very important use case for this research is for computational adept material scientists who are interested in solving specific correlation problems regarding distance lists in terms of machine learning. For example, they may be interested in solving the inverse problem directly from distance lists, or even predicting the space group of a material etc. While there is still much work to be done to fully publish this database to the public, this summer we have set up the groundwork and database maintainer code to do so.

Conclusion

This research entailed creating software and database pipelines to expedite material science researchers in the specific subfield of crystallography. While we do not have explicit findings in this research, we have created proof of concept code and preliminary results for a database of distance lists so that researchers can computer pair distribution functions on the fly. Ultimately, this research takes a valuable incremental step to what has been solved in the inverse problem so far, and we hope that with this contribution, it allows researchers in the future to solve this interesting problem in a faster and more efficient manner.

Acknowledgements

We would like to thank the Columbia SURE program to fund and make this research possible.

References:
[1] Kodama, K., Iikubo, S., Taguchi, T., & Shamoto, S. I. (2006). Finite size effects of nanoparticles on the atomic pair distribution functions. Acta Crystallographica Section A: Foundations of Crystallography, 62(6), 444-453.
[2] Billinge, S. J. (2010). The nanostructure problem. Physics, 3, 25.
[3] Juhás, P., & Billinge, S. J. L. (2017, January). DiffPy-CMI-a software toolbox for structure analysis by Complex Modeling method. In ACTA CRYSTALLOGRAPHICA A-FOUNDATION AND ADVANCES (Vol. 73, pp. A388-A388). 2 ABBEY SQ, CHESTER, CH1 2HU, ENGLAND: INT UNION CRYSTALLOGRAPHY.