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

Sean Wu, Simon J.L Billinge

Crystallography and X-ray diffraction are powerful tools for understanding the atomic composition and structure of materials. Information extracted from these techniques is crucial for material scientists as it can assist in the development of improved drugs and even more efficient battery designs. Crystallography involves two fundamental problems: the forward and inverse problems. The forward problem is relatively straightforward, as scientists can diffract X-ray beams off the material to obtain a scattering pattern, which provides interatomic information. However, the inverse problem, especially in complex nano-materials, is more challenging, and advanced methods are required to determine the 3D electron density map from one-dimensional diffraction data. The pair distribution function (PDF) is employed for this purpose, representing a weighted histogram of interatomic distances. Computing the distance lists of atoms involves a double sum, resulting in a nested for loop with $O(n^2)$ time complexity. This computational burden can significantly prolong data analysis for high r(Å) values. To address this issue, we propose a solution that involves pre-computing a distance list database. By doing so, we can calculate the pair distribution function in linear or even constant time, enabling the analysis of more complex problems. In this study, we developed a Python framework that not only builds but also maintains a distance list database. The database contains 785 experimental.CIF files and an extensive .json dataset, including metadata such as space group and unit cell information. We believe this resource will be valuable for computational material scientists, especially in the context of machine learning, where they can directly solve the inverse problem or other correlation problems from distance lists. While the database is not yet fully published, the groundwork and database maintainer code have been established during the summer to facilitate future research and applications in crystallography.