Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Interpretable crystal descriptions across length scales for materials discovery
Interim Report
Project Dates: 01/09/2020 - 28/02/2021
School of Chemistry & School of Informatics, University of Edinburgh

Dr James Cumby & Dr Sohan Seth
The University of Edinburgh

Report Date: 27/04/2021

AI3SD-Project-Series:Report[x]-Interim

Interpretable crystal descriptions across length scales for materials discovery
AI3SD-Project-Series:Report[x]-Interim
Report Date: 27/04/2021
Internal Circulation Only

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Interpretable crystal descriptions across length scales for materials discovery
Funding reference	AI3SD-FundingCall2-006
Lead Institution	University of Edinburgh
Project Dates	01/09/2020 - 28/02/2021
Website	Functional Materials Group Website
Keywords	Materials; descriptors; crystal structure; bulk modulus

2 Project Team

2.1 Principal Investigator

Name and Title	Dr James Cumby, Lecturer in Inorganic Chemistry
Employer name / University Department Name	University of Edinburgh, School of Chemistry
Work Email	james.cumby@ed.ac.uk
Website Link (if available)	Functional Materials Group Website

2.2 Co-Investigators

Name and Title	Dr Sohan Seth, Senior Data Scientist
Employer name / University Department Name	University of Edinburgh, School of Informatics
Work Email	sohan.seth@ed.ac.uk
Website Link (if available)	N/A

2.3 Researchers & Collaborators

Dr Ruizhi Zhang is the postdoctoral research associate employed by the project.

3 Publicity Summary

Most technological devices depend in some way on crystalline inorganic materials, from the perovskite oxides found in the capacitors underpinning phones and computers through to the ceramic materials used to insulate ovens and hobs. Future technologies will require new materials with different properties, but discovering these is a significant challenge; trial and error is simply too complex and time-consuming. An alternative approach is to harness our knowledge of the crystalline structure of existing materials in order to predict the properties of new ones, using machine learning (ML). Unfortunately, the conventional way in which we represent crystal structures is unsuitable for current ML methods. This project aims to develop new ways to represent structures as an input for ML, and ultimately to predict physical properties (such as how hard a material is) based on its atomic structure.

4 Executive Summary

The aim of this project is to develop two new ways of describing crystallographic extended solids to enable their effective application in machine learning pipelines. The first is a real (atomic) space descriptor based around atom-atom distances designed to capture short-range interactions, whilst the second is based on a diffraction (reciprocal) space view which captures more long-range, periodic information.

This report discusses the implementation of the real-space descriptor, and the finding that it can predict bulk modulus from crystallographic structure better than comparable approaches that do not include atomic species information. In addition, we describe the extension of this project towards a new method of quantifying crystallographic similarity using the earth mover's distance (EMD) metric.

5 Aims and Objectives

Crystallographic data (both for inorganic and organic materials) present a wealth of information that could be used for data-driven discovery, but their standard format of a periodic unit cell and fractional atomic coordinates are unsuitable due to lack of invariance; there are infinitely many unit cells to describe the structure, and permutation of atomic positions does not change the resulting structure. Although significant work has been performed to address this challenge, much of the focus has been on molecular materials due to their pharmaceutical importance. As the molecules making up these materials are inherently finite, existing methods tend to work best over short length scales or with few atomic species. In contrast, inorganic materials are effectively infinite and present a huge atomic diversity, from simple 1-atom metals to 24-atom minerals. Even for existing long-range methods, this poses a significant challenge, particularly for "small" data sets where deep learning cannot be applied.

The primary aim of this project was to develop new descriptors for extended crystal structures that could be applied across a wide range of materials chemistry problems, and are suitable for relatively small-scale machine learning approaches. The two approaches investigated can be categorised in terms of the crystal 'space' in which they operate:

Real Space This is the position of atoms within a crystal, and the space between them. The descriptor is based on the pair-wise distances between atoms, grouped by their proximity.

Reciprocal Space This is the Fourier transform of the atomic positions, as experimentally observed through *e.g.* X-ray diffraction. As the Fourier transform of an infinitely periodic arrangement of points is also infinitely periodic and point-like, this descriptor is focused on using the magnitudes of these points ('intensities') to capture long-range atomic ordering information.

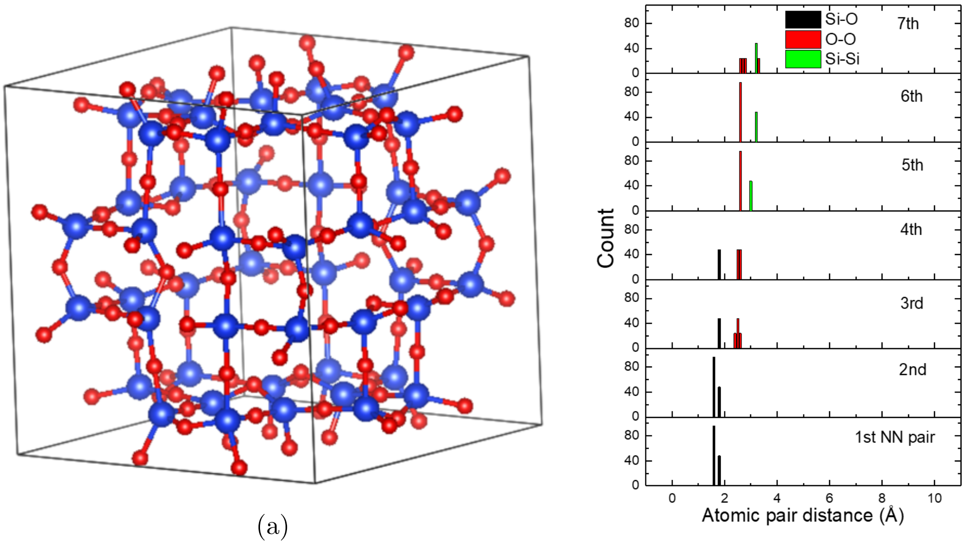
In addition to developing the new descriptors, another important objective of this project was to test their applicability and generalisation. Whilst the initial proposal aimed to apply the descriptors to the discovery of new oxyfluoride materials, more recent work has demonstrated that there are additional complexities within this area beyond the nature of descriptors used that renders it unsuitable for testing the effectiveness of new descriptors. As such, the project aimed to test the new descriptors in the prediction of physical properties such as bulk modulus from existing crystal structures. This problem has been explored previously, and gives us a benchmark to compare results with.

Following development of the real-space descriptor, it became apparent that standard distance measures (e.g. Euclidean distance) are not well-suited to comparing such distributions. As such, an additional aim was introduced early in the project to try to determine the best *similarity measures* for such descriptors, and investigate its application in existing ML approaches.

6 Methodology

6.1 Scientific Methodology

The first part of this project focused on developing an extended version of the commonly-used radial distribution function (RDF) as a real space descriptor. This is the (binned) histogram of pairwise atomic distances, up to a chosen distance cutoff. Our method takes this single histogram, and sub-divides it into multiple histograms based on the relative proximity of atoms. This introduces much more information than the standard-RDF, but the descriptor size is independent of the crystal structure complexity. The approach is best illustrated through an example, such as the binary solid SiO_2 (quartz). This material has a cubic crystallographic structure, but is relatively complex with 144 atoms per unit cell (Fig. 1a).* For each of these 144 atoms we determine their pairwise distances up to a cutoff (say 10 Å). Following this we rank the distances for each atom in ascending order, and then combine the ith entry for all 144 atoms into a single binned distribution (the ith nearest neighbour histogram, or 'ith-NN') . The resulting 'extended-RDF' (Fig. 1b) contains greater information than the standard-RDF due to the separation of equal-distance neighbours into different NN-shells, but summation across the complete range of histograms recovers the standard-RDF.



(a)

(b)

Figure 1: (a) Crystal structure of SiO_2. Si - blue spheres; O - red spheres; (b) the resulting extended-RDF.

The second (reciprocal space) descriptor calculates the X-ray diffraction intensity at points in reciprocal space using the well-known structure factor equation from crystallography,

$$F_{\mathbf{S}} = \sum_n f_{n,\mathbf{S}} \exp\left(2\pi i \mathbf{S}.\mathbf{r}_n\right)$$

*Here we have ignored crystallographic symmetry or equivalently

$$F_{hkl} = \sum_n f_{n,hkl} \exp\left(2\pi i(hx_n + ky_n + lz_n)\right).$$

Here, F is the structure factor ($F \propto \sqrt{\text{Intensity}}$), $\mathbf{S}$ is the scattering vector (*i.e.* vector between incoming and diffracted X-rays) and $\mathbf{r}_n$ (or x_n, y_n, z_n) is the fractional position of atom n within the unit cell. $h, k, l \in \mathbb{Z}$ are so-called Miller indices, and express the position of a point $\mathbf{S}$ in reciprocal space in terms of multiples of the three reciprocal lattice vectors defining the periodicity within reciprocal space (equivalent to a unit cell).[†] f_n is an atom-specific function dependent on $\mathbf{S}$ (or h, k, l) which determines how well an atom diffracts X-rays at different scattering angles. By computing F for different values of h, k, l this gives us a three-dimensional descriptor for ML applications. To overcome the problem of unit-cell invariance, we have initially focused on using the standard Niggli reduced cell for a given structure, which gives a convention for choosing lattice vectors in real space. [1] We intend to extend this to a unit-cell invariant method by utilising the relative positions of pairs of $F_{h,k,l}$ in order to add rotational invariance.

6.2 AI Methodology

Following implementation of the crystal descriptors, we aimed to test their effectiveness at predicting properties of materials based solely on crystal structure, and also investigate how they quantify similarity between different crystal structures. From an AI perspective, the first of these has made use of standard ML approaches such as kernel ridge regression (KRR) or LASSO regression, due to their readily available implementations such as scikit-learn. Models have been trained on a subset of 12,731 materials from the materials project database [2] for which calculated bulk and shear moduli are available. Both linear and radial basis function kernels have been tested, and model parameters optimised using a grid search with cross validation, using mean absolute error (MAE) as the training metric. These tests have found that a linear kernel with a regularisation strength of 1 gives the lowest MAE.

For comparison, models have also been trained using the standard RDF for comparison with our extended RDF result, using the RadialDistributionFunction method of the MatMiner package. [3]

For similarity calculations, we have explored the use of the Earth Mover Distance (EMD) in order to compute a distance metric between two histogram distributions. The benefit of this method over the linear product is that for two structures which are topologically identical but with different lattice parameters (i.e. different average bond lengths) the EMD will give a small distance which increases with lattice distortion, while the linear product will give a large distance regardless of the degree of expansion/contraction.

7 Interim Results

7.1 Bulk Modulus Prediction

Using both bulk modulus (BM) and $\log_{10}$(BM) (logBM) as the ground truth, our results show that logBM gives a better model for both training and test data (Fig. 2). This can be explained due to the skewed BM distribution, with a relatively small number of large BM materials from which to learn important parameters. Because of this skew, many previous reports predicting BM have often removed such extreme values, but we have chosen not to to maintain generality.

Comparing the MAE obtained for the logBM data with the standard RDF descriptor, it is readily apparent that the extended RDF vastly outperforms the existing descriptor for all training set sizes tested (Fig. 3). For large training sets, this approaches the current state-of-the

[†] Equivalently, h, k, l can be considered as a set of parallel planes in real space.

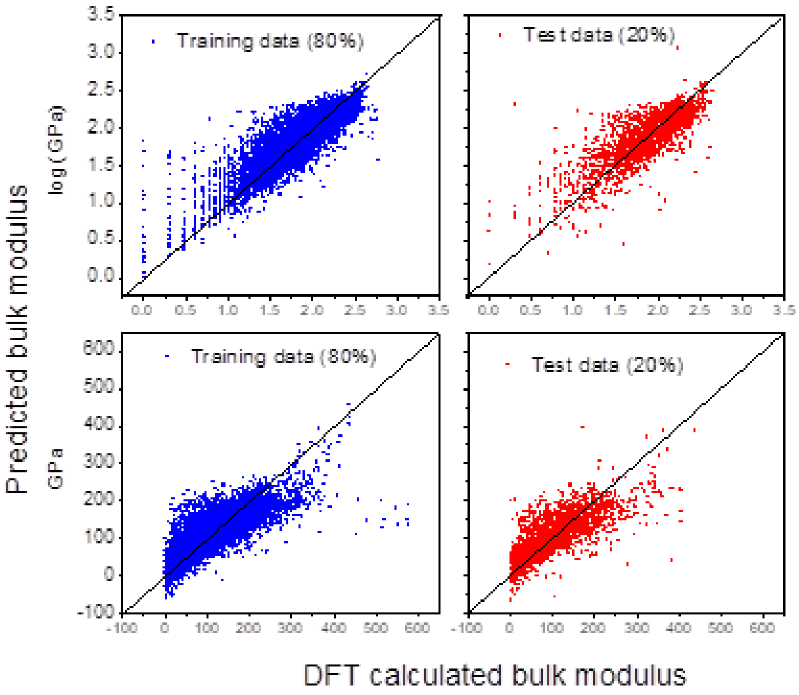


Figure 2: Predicted vs ground truth bulk modulus for training and test data for the full dataset, trained using both BM and log(BM) as target value.

art in BM prediction of MAE (log(GPa)) $\simeq 0.05$. [4] It is worth highlighting that this state of the art (based on crystal graph convolutional neural networks) is trained with a large number of descriptors, including both atom-based and structure-based information. In contrast, the extended RDF method does not include any details about atom types, yet still reproduces the computed bulk modulus value with remarkable accuracy.

We have also trained a model using bulk modulus data present in the Materials Project database as of 2017 for comparison with the work of [4] (yellow curve in Fig. 3)[‡] which reveals that older BM data give rise to slightly better predictions. This can be rationalised due to the computational expense of computing BM using density functional theory (DFT), hence initial attempts within the materials project focused on higher-symmetry crystal structures which are potentially more numerous.

We have also tested a number of different ML regression models (KRR with linear and gaussian kernels, LASSO, random forests, kNN regression) and have found that linear-kernel KRR performs most effectively for this dataset. Further work is on-going to explore alternative kernels that may be better suited to the problem.

7.2 Structural similarity

An underlying requirement of all of these regression models is the ability to compute similarity between two input crystal structures, most commonly the L_2-norm (or Euclidean distance). For our extended RDF feature, this is not an accurate measure of similarity between two crystal structures. Take, for example, the structure of a material measured at different temperatures. Thermal expansion means that each pairwise distance will change with temperature, with the result that they will no longer occupy the same histogram bins in the extended RDF. The L_2-norm will give a large distance between these two structures, comparable with that of a completely different structure with no overlapping bins, whereas arguably the thermal expansion

[‡] the exact dataset used in [4] is unknown, but this is a best approximation

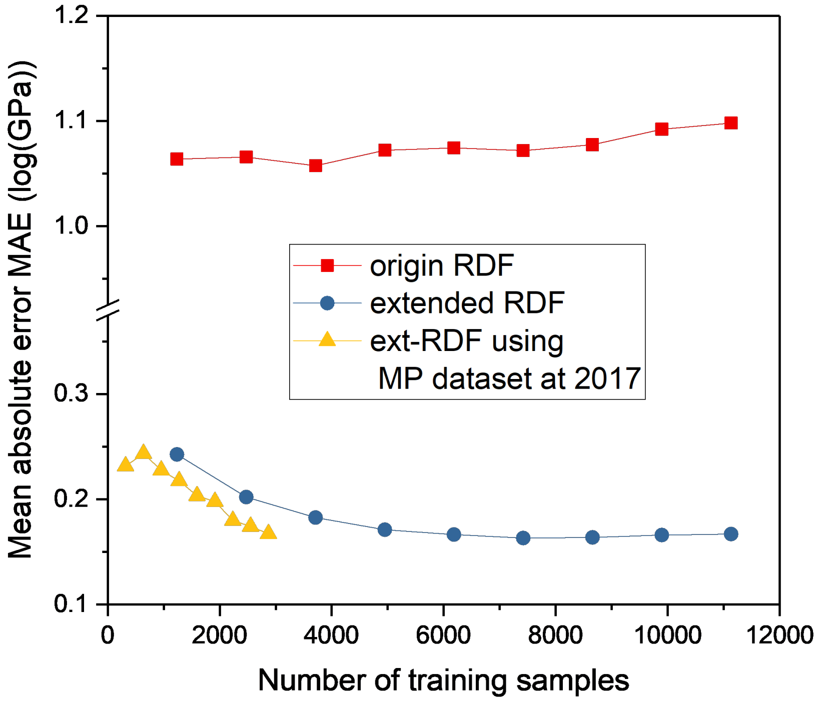
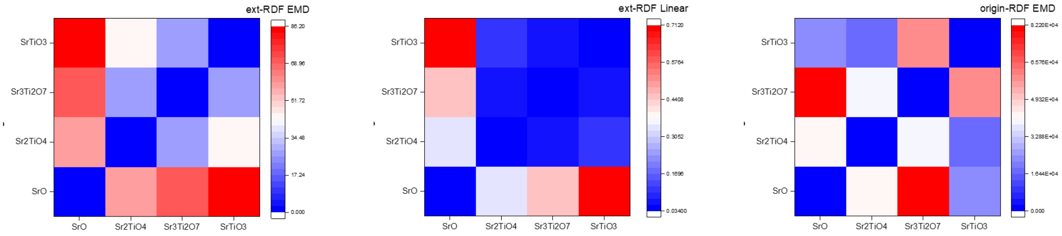


Figure 3: Variation of MAE with training set size for extended RDF, standard RDF (”origin RDF”) and a specified data subset.

should be a minor perturbation.

One solution to this problem is to employ the earth mover’s distance (EMD), which determines the minimum cost to equate two distributions. Such a method has recently been employed to calculate similarity between chemical compositions. [5] We have applied this metric to a number of crystallographically-related materials to compare it based on chemical understanding. As an example, we have considered the Ruddlesden-Popper structural series $\mathrm{Sr}_{n+1}\mathrm{Ti}_n\mathrm{O}_{3n+1}$ for $n = 0, 1, 2, \infty$. This structure can be considered as an intergrowth of rocksalt ($\mathrm{SrO}, n = 0$) and perovskite ($\mathrm{SrTiO}_3, n = \infty$) layers with varying ratios. Comparing the EMD and cosine similarity for this series (Fig. 4) it is clear that the EMD gives a measure of similarity consistent with that expected chemically, whilst the cosine similarity clearly differentiates SrO from SrTiO_3, but barely discriminates $n = \infty$ from $n = 1, 2$. Additionally, EMD applied to the standard RDF gives almost meaningless comparisons, with SrTiO_3 appearing more similar to SrO than $\mathrm{Sr}_3\mathrm{Ti}_2\mathrm{O}_7$, even though the latter contains a greater proportion of perovskite.



(a)

(b)

(c)

Figure 4: Pairwise similarity for the Ruddlesden-Popper series ($n = 0, 1, 2, \infty$) using (a) EMD on the extended RDF; (b) cosine similarity of the extended RDF and (c) EMD on the original RDF.

8 Outputs

The code generated during this project is stored in a private GitLab repository, but will be made publicly available to coincide with publication of the results of the project. The initial results have been presented at internal seminars within both the School of Informatics and School of Chemistry at the University of Edinburgh. Additionally, we anticipate presenting the results at the 25th Congress of International Union of Crystallography (IUCR) conference in Prague, 14th-22nd August 2021.

9 Progress Summary

So far, the project has implemented a new real-space descriptor and explored its applicability in predicting physical properties from crystal structure, for comparison with existing methods. In addition, we have implemented the Earth Mover's Distance (EMD) as an alternative distance measure for comparing crystal structures. Both objectives have shown promising results, and work is on-going to combine them.

We have also implemented an initial approach to the reciprocal space descriptor and are performing preliminary analyses; further results will be published in our final project report.

10 Next Steps

Initially, we plan to continue implementing and testing the reciprocal space descriptor, and then combining it with the extended RDF method to provide more accurate predictive models. We also intend to incorporate the EMD results where possible. In the longer term, we plan to document and make publicly available our newly developed code. We expect that the results of this project will produce at least two publications in peer reviewed journals; preparation of the first (on the extended-RDF method) is currently underway.

11 References

References

[1] R. W. Grosse-Kunstleve, N. K. Sauter, P. D. Adams, *Acta Crystallographica Section A Foundations of Crystallography*, 2003, **60**, 1–6.

[2] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Materials*, 2013, **1**, 011002.

[3] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Computational Materials Science*, 2018, **152**, 60–69.

[4] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chemistry of Materials*, 2019, **31**, 3564–3572.

[5] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, M. J. Rosseinsky, *Chemistry of Materials*, 2020, **32**, 10610–10620.

12 Data & Software Links

N/A