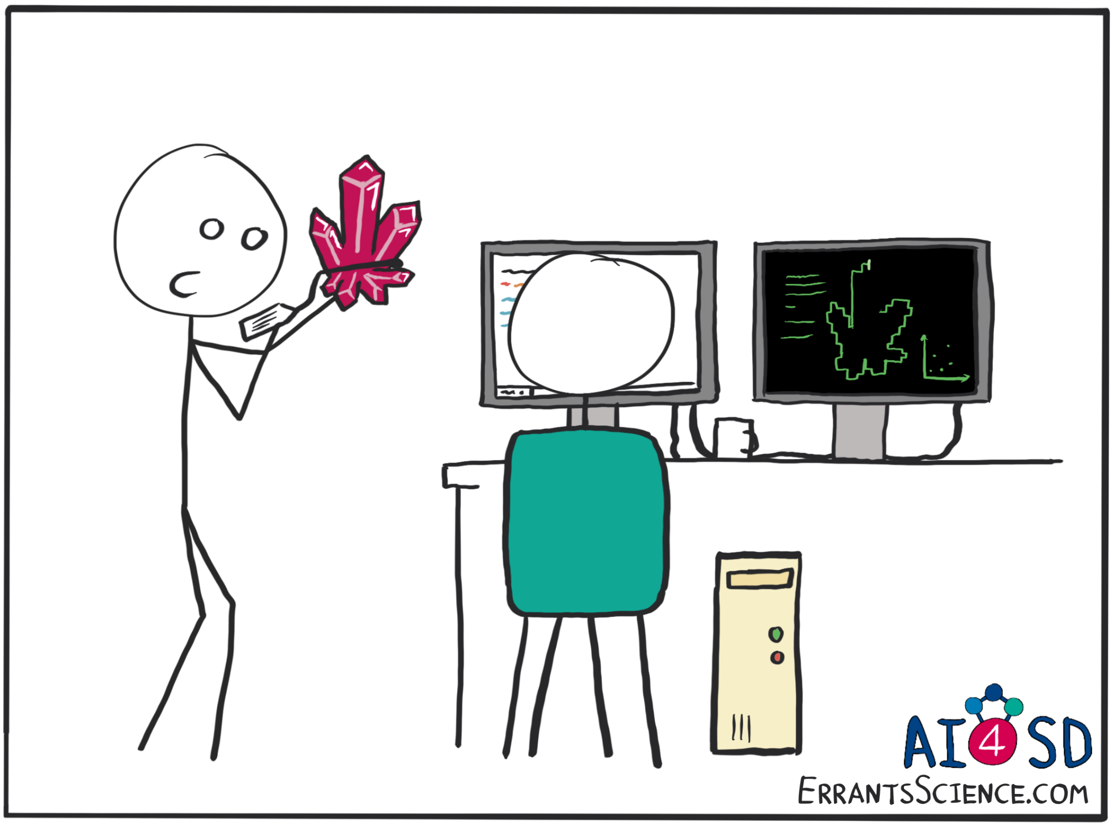
AI 4 SCIENTIFIC DISCOVERY

AI 4 Science Discovery Network+

Interpretable crystal descriptions across length scales for materials discovery Final Report
Project Dates: 07/09/2020 - 01/03/2021
School of Chemistry & School of Informatics, The University of Edinburgh



Dr James Cumby & Dr Sohan Seth
The University of Edinburgh

Report Date: 16/07/2021

AI4SD-Project-Series:Report7_Cumby_Final

Interpretable crystal descriptions across length scales for materials discovery
AI4SD-Project-Series:Report7_Cumby_Final
Report Date: 16/07/2021
DOI: 10.5258/SOTON/P0038

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	AI4SD-FundingCall2-006
Lead Institution	The University of Edinburgh
Project Dates	07/09/2020 - 01/03/2021
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	The 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	School of Informatics, University of Edinburgh
Work Email	sohan.seth@ed.ac.uk
Website Link (if available)	Personal website

2.3 Researchers & Collaborators

Dr Ruizhi Zhang was 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 might not be suitable 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 atomic structure.

4 Executive Summary

The aim of this project is to develop new ways of describing crystallographic inorganic solids to enable their effective application in machine learning pipelines. The majority of this work has focused on an approach to group interatomic distances into histograms (referred to as GRID), with the benefit that it is easily (and quickly) calculated, and readily interpreted. Coupled with this distance-based descriptor, we have also investigated the use of earth mover's distance (EMD) as a dissimilarity metric between the respective histograms.

Our results show that the GRID descriptor is able to quantify similarity between crystal structures in a similar way to a chemist, particularly when combined with EMD. This is in contrast to existing distance-based descriptors, which do not reproduce chemical intuition. Furthermore, we have demonstrated the GRID+EMD, combined with a simple k-nearest neighbours model, is able to accurately predict bulk moduli from crystal structures with an accuracy approaching the state-of-the-art.

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 an unsuitable representation due to lack of invariance; there are infinitely many unit cells to describe the structure, and permutation of atomic labels 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 methods such as graph convolutional neural networks may not 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, collected into groups based on 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.

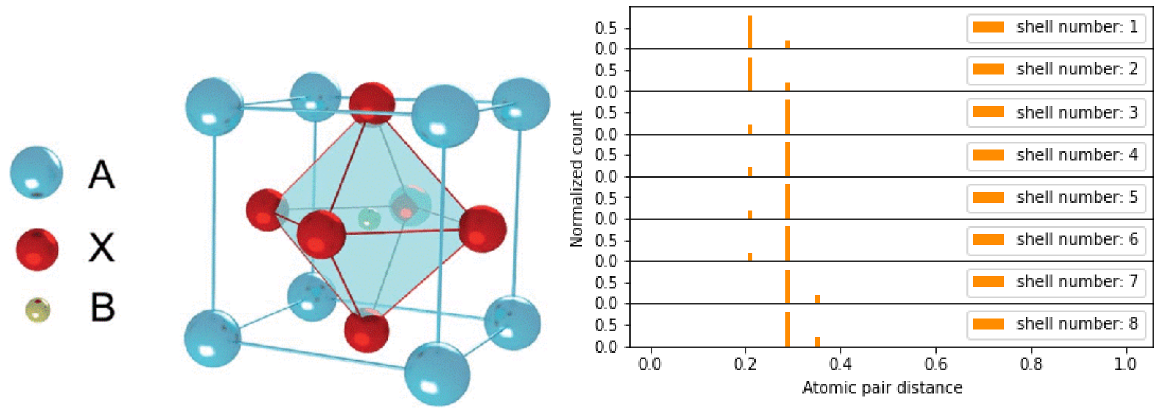
In addition to developing the new descriptors, another important objective of this project was to test their applicability and generalizability. Whilst the initial proposal aimed to apply the descriptors to the discovery of new oxyfluoride materials, more recent work has demonstrated that there are additional complexities within this area beyond the choice of descriptors that renders it unsuitable for testing the effectiveness of the project aims. As such, new descriptors generated during the project were tested for 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; the grouped representation of interatomic distances (GRID). GRID introduces much more information than the standard-RDF, but importantly the descriptor size is independent of the number of atomic species. The approach is best illustrated through an example, such as the archetypal perovskite structure ABX_3 (e.g. $SrTiO_3$) with five atoms per unit cell (Fig. 1a). For each of these unique 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 atoms into a single binned distribution (the ith nearest neighbour histogram, or 'ith-group'). The resulting GRID (Fig. 1b) contains greater information than the standard-RDF due to the separation of equal-distance neighbours into different NN-groups, but summation across the complete range of histograms recovers the standard-RDF.



(a)

(b)

Figure 1: (a) Crystal structure of ABX_3. (b) the first eight groups of the resulting GRID.

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,

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

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 as 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, giving a convention for choosing lattice vectors in real space.[1] We aim 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 GRID result, using the RadialDistributionFunction method of the MatMiner package.[3]

A significant challenge encountered during this project was the difficulties in accurately quantifying similarity between two histograms such as GRID. For example, expanding the unit cell slightly (as in thermal expansion) results in bond distances falling in to different histogram bins, but clearly the similarity between two structures should be proportional to the amount of expansion. In contrast, “standard” measures of dissimilarity such as Euclidean distance or cosine similarity are not well suited to changes in binned distributions; as an example, the vector [0,1,1,0,0,0] has the same Euclidean distance to [1,0,1,0,0,0] as [0,0,0,0,1,1]. To overcome this, we have adopted the Earth mover’s distance (EMD) to compute dissimilarity between sets of histograms, which considers how much ‘effort’ is required to transform one distribution into another. Although relatively uncommon in machine learning literature, such an approach has very recently been employed to measure compositional similarity in crystals. [4]

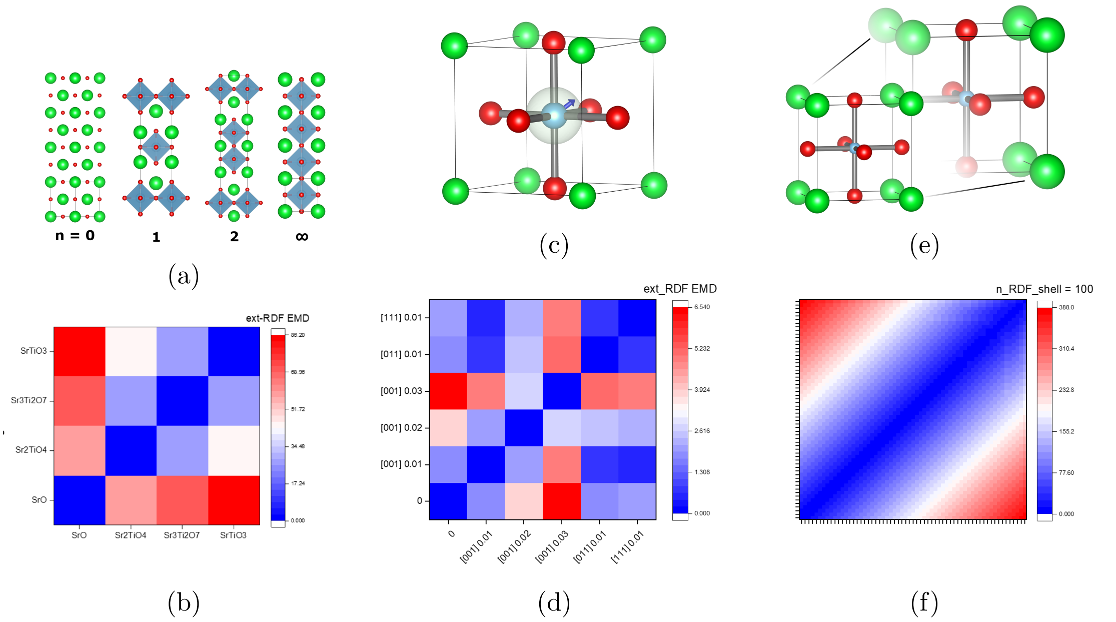
7 Results

7.1 Structural similarity

In order to test the efficacy of GRID combined with EMD in quantifying structural similarity, we have constructed a number of artificial examples based on known crystal structures. These were designed to address the effect of long-range structural modifications (layered Ruddlesden-Popper structures, Fig. 2a), short-range atomic displacements resulting in subtle symmetry

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

changes ($BaTiO_3$, Fig. 2c) and the effect of lattice expansions (perovskite, Fig. 2e). In all cases the results agree with chemical intuition. In the case of the layered Ruddlesden-Popper ($Sr_{n+1}Ti_nO_{3n+1}$ for $n = 0, 1, 2, \infty$) phases, GRID+EMD finds that as more perovskite layers are inserted between rock-salt SrO layers the structures become quantitatively more similar to the $n = \infty$ phase. In contrast, combining GRID with cosine dissimilarity gives less distinction between the intermediate phases, whilst the standard RDF (with either EMD or cosine dissimilarity) results in greater similarity between SrO ($n = 0$) and $SrTiO_3$ ($n = \infty$) than with some of the intermediate compositions. For atomic displacements within $BaTiO_3$ GRID is proportional to the degree of displacement, as expected. In the final example (lattice expansion of a cubic ABX_3 perovskite) GRID+EMD gives a continuous increase in dissimilarity with isotropic expansion, in stark contrast to discontinuous behaviour observed using the original RDF.



(a) (c) (e)

(b) (d) (f)

Figure 2: Pairwise Earth mover distances for (a, b) the Ruddlesden-Popper $Sr_{n+1}Ti_nO_{3n+1}$ series ($n = 0, 1, 2, \infty$); (c, d) Ti displacements within $BaTiO_3$; and (e, f) lattice expansions within cubic perovskite ABX_3.

7.2 Property prediction

Using the newly developed GRID+EMD approach, we have tested it with a number of machine learning models in its ability to predict bulk modulus from atomic structure (Fig. 3). Using the simple k-nearest neighbours (kNN) approach with $k = 1$, we find that we are able to predict bulk modulus with a mean absolute error (MAE) of 18 GPa when incorporating both structural (GRID) and composition information, approaching the state-of-the-art value of 10 GPa. The median absolute error (arguably a better measure due to the skewed bulk modulus distribution, which previous studies have overcome by removing 'outliers') is 11 GPa. It is noteworthy that the state-of-the-art uses complex graph-based convolutional neural networks (GCNNs), while our approach simply interpolates to the most similar known material.

7.3 Fourier space description

We have implemented a method to compute X-ray diffraction intensities on a h, k, l grid from crystal structures, and have used the same bulk modulus data with a KRR model to test its

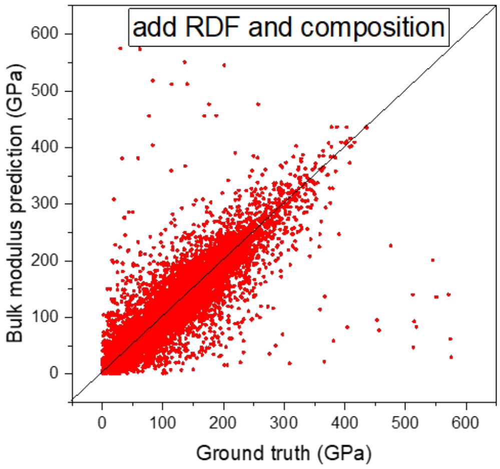


Figure 3: Distribution of predicted and ground-truth bulk modulus values, showing that most predictions lie close to the ideal x = y line, with a few significant outliers resulting in higher MAE overall.

applicability as a descriptor. Visualising the distribution of predicted and true bulk moduli (Fig. 4) it is clear that while the trained model gives reasonably accurate agreement between predicted and ground-truth values, the model does not generalise well to the testing data (MAE = 28%). We attribute this to the discontinuous unit cell changes between structurally similar materials, such as changes in symmetry due to small atomic displacements; work is continuing beyond the end of this project to extend the implemented code to a unit-cell-independent model, including a AI4SD-funded summer internship position investigating the link between Fourier and real space descriptors.

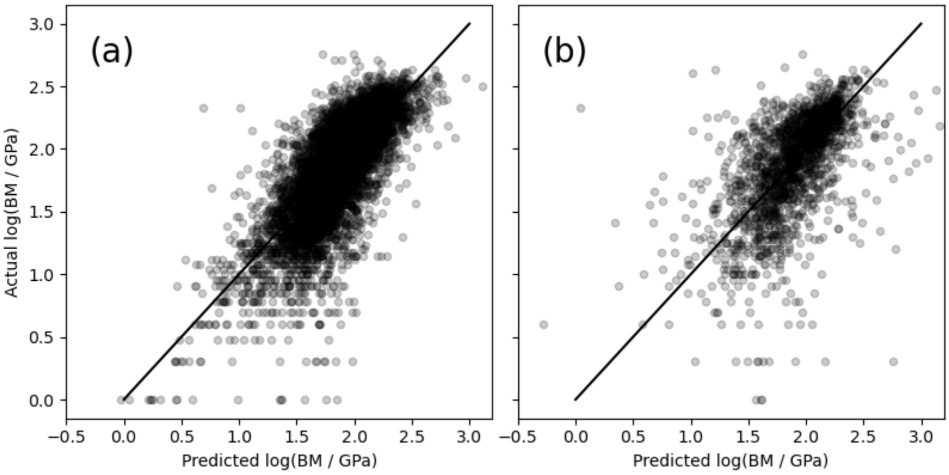


Figure 4: Distribution of ground truth and KRR-predicted bulk moduli (on a log(BM / GPa) scale) using the Fourier space descriptor for (a) training and (b) test data (80:20 split).

8 Outputs

Publications

- R. Zhang, S. Seth and J. Cumby, Grouped representation of interatomic distances (GRID) as a similarity measure for crystals, *in preparation*

Presentations

- J. Cumby, Machine learning for materials and chemicals (AI4SD Summer school), Online, August 2021.
- J. Cumby, 25th Congress of International Union of Crystallography (IUCR) conference, Prague, August 2021.
- J. Cumby, Scottish Dalton meeting, Online, August 2021.

Code

The code generated during this project will be made publicly available prior to publication at the Functional Materials Group GitLab site.

9 Conclusions

We have developed a new descriptor based on interatomic distances which shows promise in predicting physical properties from inorganic crystal structures with accuracy approaching the current state-of-the-art. Importantly, we have demonstrated the effectiveness of the earth mover's distance in comparing these distributions, and implemented code to apply it to different problems. This project has also allowed us to explore different technical aspects of applying ML to crystal structures, such as considering a wide range of kernel functions, distance metrics and ML models. In addition to resulting in publications, these explorations will have much broader impacts, and will be invaluable in directing and evaluating future work within the research group.

We have also implemented code to generate Fourier-space descriptions of crystal structures, and performed preliminary testing on its use as a descriptor. Due to the additional need to investigate alternative distance metrics progress on this Fourier-space descriptor was slower than expected, but the code will be further developed and explored.

10 Future Plans

This project has strengthened the existing collaboration between JC and SS, and enabled the generation of many previously unforeseen research goals. As such, it will form the basis of future research grants to explore the developed descriptors further, and to apply them to cutting-edge materials problems. In the short term, the Fourier-space code developed will form the basis of an AI4SD-funded summer internship investigating the link between real- and reciprocal-space descriptors using deep learning. Beyond this, we anticipate applying for more substantial research grants to develop (and disseminate) the methods further.

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

11 Data & Software Links

Code will be published on the Functional Materials Group GitLab site associated with the relevant publications.