Crystal Structure Prediction via Semi-supervised Clustering

Project Plan

Tom Kuipers

CS4821 - MSci Final Year Project

Supervised By: Dr Argyrios Deligkas

Department of Computer Science

Royal Holloway, University of London

1 Abstract

Crystal structures can take on many forms and have many interesting properties. One well-known example of a crystal is *sodium chloride* (NaCl), used by the human body to regulate fluid levels. Another example whose use was prolific in the 1970s and still is today, is quartz or rather silicon dioxide (SiO_2). This crystal is used in many time-keeping devices due to its piezoelectric properties.

In a crude sense, crystals are created by combining a set of given elements together, giving a so-called composition. Different types of crystal structures can be yielded depending on the type of elements combined. For example, NaCl is an ionic crystal — it combines a metallic (Na^{+1}) ion with a non-metallic (Cl^{-1}) ion. However, a given composition of a crystal can only exist if the sum of the charges of all ions is zero.

Put formally, crystals are solid materials in which the atoms are arranged in a highly ordered configuration. Within such a material, there always exists a local configuration of atoms, so-called a *unit cell* which is of low relative energy. It is from this cell an entire crystal can be synthesised. Crystal structures are defined by tessellating the unit cell, preserving *periodicity* in all axis. Whilst these are well-established crystals, for many other potential structures it is not immediately clear whether a stable configuration exists.

To establish whether a stable configuration is likely to exist, in theory we could do the following: given an empirical formula, $SrTiO_3$ for example and a 3D grid in which to place 10 atoms of each element, we can discretise the grid to $20 \times 20 \times 20$. We then place the atoms in the grid in some configuration, yielding a crystal structure with a relative energy value. This energy value then defines a point on the *potential energy surface*. The structure with the lowest energy correlates to a potentially stable crystal structure which may then be able to be synthesized.

The issue with the method above is that we have $\binom{20^3}{50} \approx 4.025 * 10^{130}$ possible combinations of atom placements. This is an over-estimation as some configurations may lead to atoms overlapping however it suffices to show that in general, it is not feasible to test all possible configurations to obtain the minimum energy structure. In addition, this method is exponential with respect to the number ions and elements involved. However, if we can correctly and efficiently predict the structure of unknown crystals, we can accelerate the discovery for new materials, potentially those with uses in the pharmaceutical industry, for example. This forms the basis of the Crystal Structure Prediction problem, which has been a discussion in physical science since the 1950s [3]. For example, CSP lead to the discovery of polymorphs of progesterone, a steroid used to treat abnormalities in menstrual cycles [7].

Computationally, the CSP problem is formulated as an optimisation problem and several approaches have already been introduced to tackle this problem such as basin hopping [5], simulated annealing [6], and evolutionary and genetic algorithms [2]. An approach of interest in this project is the Flexible Unit Structure Engine (FUSE) [1] which is based on Monte Carlo basin hopping. In essence, starting from an empirical formula, it begins with building blocks, so-called *motifs* containing a subset of atoms in some configuration which are then combined to form a unit cell/*probe structure*. There are then six potential permutations of the unit cell which can be selected to improve the relative energy of the structure produced in each evolution.

Whilst these methods have seen some success in predicting structures more efficiently, many methods focus on improving the underlying search problems as opposed to understanding the relationship between permutations of the atom point set and the point on the potential energy surface. Here is where I believe machine learning can aid in giving further insight and understanding to the CSP problem and this is the main focus of my project.

Machine learning approaches do already exist such as Ryan, Lengyel and Shatruk's [8] method which utilises a *deep neural network* (DNN). The method analyses known crystal structures, distinguishes chemical elements from their crystallographic environment and identifies structurally similar atomic sites. In doing so, it aims to discover templates for atomic sites in which atoms could then be placed combinatorially and predict the likelihood that a new compound can be formed.

In a machine learning context, the modelling of the crystallographic data is crucial for any algorithm to be able to extract the desirable relationships. In general, fingerprints are used to model chemical data in purely numerical contexts. In the CSP context, a fingerprint represents some kind of patterns unique to various crystal structures. In the aforementioned deep learning paper, the input is defined as an atomic fingerprint (AFP) which represents the geometric relations between the atoms within a crystal structure [8]. Other fingerprints exist such as Valle and Oganov's Crystal Structure Fingerprint (CSFP) [9], however, the one of interest in this project is the Density Fingerprint in which crystal structures are modelled as periodic point sets and we are interested in how close given structures are to being isometric [4], irrespective of atom species.

With this project, I aim to bring together various ideas from the aforementioned works and formulate a species-conscious algorithm that will perform semi-supervised clustering on periodic point sets in order to evaluate whether candidate crystal structures seem promising and warrant further evaluation. The main goal is to produce an algorithm that can cluster potential crystal structures into classes depending on their relative energy. In doing so, the hope is to develop an understanding and discover relationships between the point sets, atom species and points on the potential energy surface. From there, I aim to produce an engine which chemists can use that, given a set of block motifs (similar to those used in FUSE [1]), can efficiently explore potential structures in specific classes of crystal structures, presenting those most likely to be stable. A secondary goal is to then perform geometry optimisation on these structures in order to yield the minimum energy unit cells for potential crystal structures.

2 Timeline

Ideally, my plan will be to focus on implementation during term one and focus on the analysis, fine-tuning and evaluation in term two. Should the plan go smoothly, I will also leave some time to further augment my ideas.

2.1 Term 1

- Week 1: Study Density Fingerprint papers.
- Week 2: Augment and formulate method to handle ion species.
- Week 3-4: Implement the code for representing data in ML context.
- Week 5-6: Create data points in high-dimensional space and convert existing data into a usable and uniform standard.
- Week 7: Research and evaluate clustering methods.
- Week 8-9: Fine-tune and optimise model parameters.
- Week 10-11: Prepare for interim report and presentation.

2.2 Term 2

- Week 1-2: Augment the approach further and re-evaluate methods.
- Week 3: Implement visualisations of clusters.
- Week 4-5: Encapsulate code into an engine using proper SE practices.
- Week 6: Implement interfaces to common chemistry programs such as VESTA.
- Week 7-9: Re-evaluate final project results and extend to relaxations, time-permitting.
- Week 10-11: Prepare for final report and viva.

3 Risks and Mitigations

As with any project, there are always risks; some unavoidable. My project is no exception and although managable, it is ambitious and comes with specific risks. In this section I will start by discussing the general risks and their mitigations and then move on to those specific to my project.

3.1 Hardware Failure

It is possible for the main hardware used whilst undertaking my project to fail, potentially involving significant data loss and thus progress. It cannot be predicted when various storage media will fail however I will mitigate the risk of data loss by keeping all my code under version control, using the GitHub platform and ensure that I commit my code regularly. In addition to the code, I will also keep my report under version control using Overleaf. These measures mean that all work is stored in the cloud and is thus unaffected by local hardware failure.

3.2 Poor Estimation of Tasks

Realistic time management is crucial if a project is to succeed within a given timeframe, otherwise tasks can overrun and the overall progress can be stifled. With respect to my project, I will ensure that I divide all tasks into appropriate sub-tasks and research these tasks substantially to better understand how long they can be expected to take.

3.3 Uneven Balance Between Report/Code

The code is just as important as the reports and one can hinder the other. Becoming caught up on the code could mean that the report lags behind and vice versa. To mitigate the risk of this occuring, I will focus on the code which I estimate will take significant attention to execute successfully and allow a couple of buffer weeks at the end of each term to wrap up aspects of the reports.

3.4 Machine Learning Risks

With machine learning methods, it is possible to become trapped in a cycle of pursuing the wrong avenue without realising it and then trying to optimise the problem to fit the model and not the other way around. This wastes valuable time and risks producing methods which do not fit the problem or benefit the end goal. In this project, I will combat this by continually evaluating and scrutinizing the method at every opportunity, as well as ensuring I understand the data/output before proceeding. I will also compare my results with those of other ML approaches.

3.5 Overhead of Chemistry Knowledge

As a computer scientist, I understand the computational problem at hand however there are various aspects that are intertwined with chemistry and knowledge thereof may be crucial to understand how to proceed with my project or how to evaluate my results. In order to ensure that this does not slow me down or hinder the project in producing usable results by chemists, I will allow time for research and understanding of both the chemical and computational aspects of the project. Should any particular issues arise, I will consult with contacts in the chemistry department of Liverpool university — a department which is active in the research of the CSP problem.

3.6 Computational Efficiency

It can be the case that certain functions or solutions to sub-problems are not coded in the most efficient way possible from the beginning. If my project is to contain inefficient code, I risk ending up with a project which adds little value to the CSP problem as I may not have aided to "acceralate" the

search for new materials at all. To ensure this is very unlikely to happen, I will endeavour to write efficient code from the beginning. I will do this by evaluating the theoretical running time/estimating the asymptotic bounds of my code, as well as measuring practical running time/baselines of the code throughout the project.

3.7 Conscious Experiments

Experimenting with machine learning can be quite a time-consuming task, especially depending on the amount of data involved. It is possible for experiments to be run that ultimately add little to what is already known or are wasted due to lack of robust code. In this project, I will write all code using Test-driven Development and use industry-standard software engineering practices to avoid such issues. I will also evaluate the results of my code at regular intervals to ensure that I tweak my model and hyperparameters before wasting time.

Acronyms

CSP Crystal Structure Prediction.

DNN Deep Neural Network.

FUSE Flexible Unit Structure Engine.

ML Machine Learning.

SE Software Engineering.

TDD Test-driven Development.

Glossary

GitHub A cloud-hosted extension of a Git version control system with a web GUI.

Overleaf An online LaTeX editor providing features such as history/version control and collaboration.

VESTA A 3D visualization program for structural models, volumetric data such as electron/nuclear densities, and crystal morphologies.

References

[1] C. Collins, G. R. Darling, and M. J. Rosseinsky. 2018. The Flexible Unit Structure Engine (FUSE) for probe structure-based composition prediction. *Faraday Discuss.* 211 (2018), 117–131. Issue 0. https://doi.org/10.1039/C8FD00045J

[2] David M Deaven and Kai-Ming Ho. 1995. Molecular geometry optimization with a genetic algorithm. *Physical review letters* 75, 2 (1995), 288.

[3] Gautam R. Desiraju. 2002. Cryptic crystallography. *Nature Materials* 1, 32 (2002), 77–79. https://doi.org/10.1038/nmat726

[4] Herbert Edelsbrunner, Teresa Heiss, Vitaliy Kurlin, Philip Smith, and Mathijs Wintraecken. 2021. The Density Fingerprint of a Periodic Point Set. In *37th International Symposium on Computational Geometry (SoCG 2021) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 189)*, Kevin Buchin and Éric Colin de Verdière (Eds.). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 32:1–32:16. https://doi.org/10.4230/LIPIcs.SoCG.2021.32

[5] Stefan Goedecker. 2004. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *The Journal of chemical physics* 120, 21 (2004), 9911–9917.

[6] J Pannetier, J Bassas-Alsina, J Rodriguez-Carvajal, and V Caignaert. 1990. Prediction of crystal structures from crystal chemistry rules by simulated annealing. *Nature* 346, 6282 (1990), 343.

[7] R.S. Payne, R.J. Roberts, R.C. Rowe, and R. Docherty. 1999. Examples of successful crystal structure prediction: polymorphs of primidone and progesterone. *International Journal of Pharmaceutics* 177, 2 (1999), 231–245. https://doi.org/10.1016/S0378-5173(98)00348-2

[8] Kevin Ryan, Jeff Lengyel, and Michael Shatruk. 2018. Crystal Structure Prediction via Deep Learning. *Journal of the American Chemical Society* 140, 32 (2018), 10158–10168. https://doi.org/10.1021/jacs.8b03913

[9] Mario Valle and Artem R. Oganov. 2010. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallographica Section A* 66, 5 (Sep 2010), 507–517. https://doi.org/10.1107/S0108767310026395