Dissertation
Project Proposal

Project Title:
A low-dimensional map of high-dimensional data

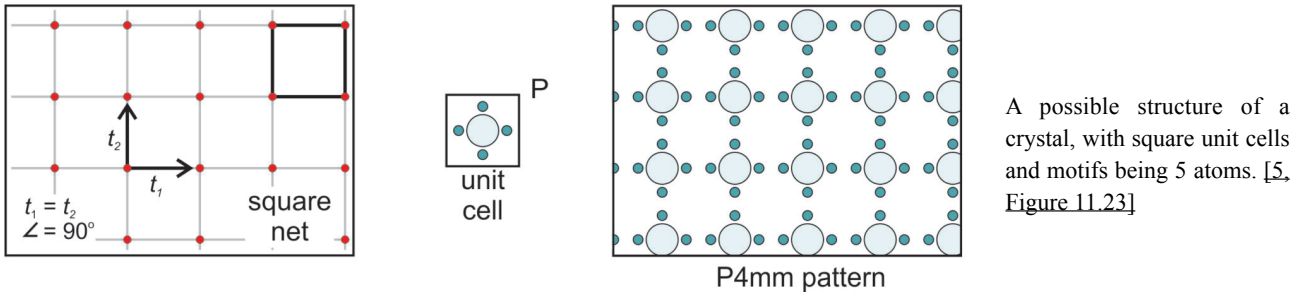
1. Project Description

This project will focus on predicting atom types given k nearest neighbor distances of the atom in some crystal structure using neural networks, in which k is a hyperparameter. The data for this project will be from real-world crystals.

Background Information

• Crystallography

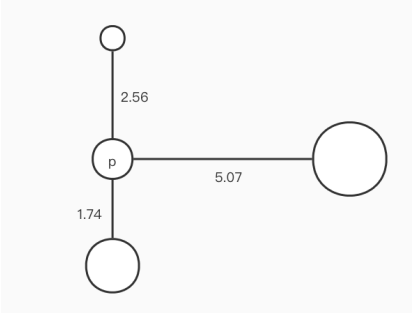
Crystals are matters with repeating arrangement of atoms, which are represented as points of different types in the 3-dimensional space. A lattice, being an infinite arrangement of points in the space, is defined by a linear combinations of a basis whose vectors span a parallelepiped called unit cell, which is the smallest building block of a crystal. The atom pattern in the unit cell is called motif, which is repeated in every unit cell. The crystal structure is then represented by a periodic point set, being the Minkowski sum of a lattice and a motif [1, Section 1]. The data used in this project will be calculated based on this model.



A possible structure of a crystal, with square unit cells and motifs being 5 atoms. [5, Figure 11.23]

• k Nearest Neighbor distances (k-NN distances)

Suppose the particular atom we are interested in is p. S is the increasingly ordered set of the distances from all other atoms in the crystal to p. k-NN distances is then defined as a vector containing the first k elements in the set S.



For example, as shown in the diagram in the left, there are in total 4 atoms, with one of them, p, being the atom we are interested in. First, we sort the distances of the atom p to all other atoms in an ascending order: S=[1.74, 2.56, 5.07]. Then the k shortest distances of the atom p is the vector consisted of the first k elements in the set S. When k=2, the 2 shortest distances of atom p is [1.74, 2.56].

2. Aims & Objectives

2.1 Aims

The aim of this project is to design a general neural network model that is capable of predicting atom types from k nearest neighbor distances after being trained. In addition, an algorithm shall be developed for a trained model to adapt to a larger k.

The possibility of developing an algorithm that could help the model to adapt to an expansion of the output set of chemical element types was also considered. However, expanding the output set is very different from expanding the input set, which is increasing the value of k. Increasing k is essentially an addition of information, and the information added tend to be less relevant, since the new atom is further from the target atom. In this sense, the trained parameters may only need a slight modification, as they have been trained to achieve acceptable performance, and they are of larger weight towards the prediction. On the other hand, the parameters on the new links need to be trained and thus increase the overall performance. In comparison, adding new chemical elements for output is completely different. As the output prediction vector is a probability distribution, when adding a new element, since the network has never learnt about it before, the output can be very random. As we train the model in order to achieve better performance on this element type, all parameters may be changed, and therefore the parameters we trained before need to be modified again to obtain similar performance on previous output set. Hence, the model should in advance include all target element types before training.

2.2 Objectives

- To develop a neural network model that is capable of predicting chemical elements given their k-NN distances.

- To develop an algorithm for a trained model to adapt to a larger k.

- To develop a method to evaluate the improvement of performance and the amplification of complexity of the model by increasing k.

  This evaluation method is required to compare the performance of different k's and suggest an optimal value.

- To suggest a best k value of complexity-accuracy tradeoff for the designed model.

3. Key Literature & Background Reading

The background reading for this project should be categorized into 3 sections:

- **Literature on Neural Network models**

  I have been learning deep learning fundamentals through *Dive into Deep Learning* [2], which provided both mathematical fundamentals and practical implementations for neural networks, and will still be a reference in the project.

  There are abundant resources providing overview of deep learning methodologies. Chapters 5-6 in *Deep Learning: Methods and Applications* [7] introduced 2 popular categories of deep networks for supervised learning (i.e. machine learning using datasets consist of labelled samples), and techniques for parameter initialization. Schmidhuber, J. (2015) [8] conducted a historical survey and summarized a massive list of important neural network models and techniques in deep learning. The section of supervised network would help to build a basic knowledge of modern successful models, and develop a list of possible models for this project.

  Particularly, Recurrent Neural Networks (RNN) based on Rumelhart, D. E., Hinton, G. E. and Williams, R. J.'s work [6] should possibly be a helpful model for this project, since it better handles sequential information, as the ordered distance vector that have different weight in this project.

- **Literature on crystal structures and related chemistry**

  As the project would attempt to design a reasonable model, chemical knowledge, especially on factors influencing the arrangement of atoms shall be reviewed.

  Introduction to Solid State Physics [10] is a classic textbook in solid state physics, the study of solids. The first few chapters of this book introduced crystal structures and the van der Waals force that provides the crystal binding. Based on the research I have done by now, the main interest of this project, which are the distances between atoms in a crystal seem to have a strong relationship with the particular chemical bond formed between them. According to Pauling, L. [11], the bond length, which is defined as the distance between two bonded atoms in a molecule, is relatively independent of the rest of the molecule. Bond length is influenced by the bond order, bond strength and the bond dissociation energy. In addition, there have been research and data analysis on this topic. For instance, Cordero, B. et al. [12] conducted a statistical analysis and deduced covalent atomic radii, which are defined as a measure of the size of an atom that forms one part of a covalent bond, for most of the elements with atomic numbers up to 96 from the Cambridge Structural Database [3]. The proposed radii showed a well-behaved periodic

dependence as expected. This explains that the distances between atoms are indeed related with the element type, which is also the theoretic foundation for the prediction model of this project to be built.

- **Literature on related previous research**

There are similar research on predicting element types based on the topology of their crystallographic environment. For example, Ryan, K., Lengyel, J. and Shatruk, M. (2018) [9] encoded the topological environment of the atom as atomic fingerprints (AFPs), before fed into a 42-layer convolutional neural network (CNN) variational autoencoder (VAE) model for training. The output was then fed into a 5-layer softmax classifier with 118 output neurons corresponding to 118 known chemical elements. The model culminated in having an average error rate of 31% on validation set. More research on this topic should be reviewed to learn designing, training and improving techniques of the model.

4. Development & Implementation

This project would use Python as the programming language, and use NumPy and TensorFlow as data processing and model training software libraries. NumPy is a library that provides tools to process high-dimensional data, while TensorFlow is a library that offers functionalities of building and training machine learning models. However, these environment choices are subject to change based on future project developments.

There are 3 reasons for choosing this development environment: First of all, I have some experiences in using this environment for neural networks construction. This helps reduce the cost for learning a new development kit. Secondly, the member from Dr. Kurlin's research group assigned to provide assistance on this project is using this environment setup. Using same development environments helps facilitate communications and cooperations with group members. Lastly, based on the literature I have reviewed, most of the relevant research have been developed in similar environments. This would bring more efficient reproduction, experiments and modification of previous models.

The project will start from shallow neural networks (i.e. neural networks with very few layers), or with an imitation of models developed in previous research. Then, updates or redesign will be made, with knowledge and techniques gained by reading the literature mentioned in Key Literature & Background Reading section.

The neural network models will be trained on GPUs and other devices provided by the University of Liverpool. The reasons will be explained in the next section.

5. Data Sources

The original data source for this project would be Cambridge Structural Database (CSD), which is a repository of three-dimensional structural data of molecules and crystals.

The data are obtained experimentally and submitted by chemists worldwide, and are freely accessible through the CCDC's website [3].

However, the particular data (i.e. k-NN distances) will be calculated from the original CSD data by the research group led by Dr. Vitaliy Kurlin [4], before used as dataset of this project. Although the original data source is freely accessible, the distances data are extracted, cleaned and processed by group members using the algorithms they developed. Therefore, to protect group members' work, those data shall not be leaked to anyone unrelated, and shall only be processed on my personal devices or the facilities provided by the University of Liverpool.

6. Testing & Evaluation

There are 3 main testing criteria of the project: the overall performance of the final model; the efficiency of the algorithm that handles the increase of k; the optimality of the value of k selected.

Based on the 3 criteria above, 3 primitive ideas of testing plan are proposed:

- The precision and recall of the model. (The performance of the model on training set, testing set and validation set).

  During this stage, the performance of the model on perturbation (a slight shift or change of the arrangement of the atoms in the crystal) was also considered. However after review ing and consulting with the supervisor, as explained before in the Key Literatures & Background Reading section, since the atom spacing is highly dependent on the chemical bonds they form, the distances would remain stable over perturbations.

- Rounds needed to improve the performance over an increase of k

  Upon an increase of k, the model have to adjust trained parameters on old links, and train all the parameters on new links. After training, the model shall get an improved performance, or at least similar performance, because the increasing of k, as stated before, is essentially an addition of information. Therefore, the rounds of training needed for the model to have an improved performance approximatively describes the time efficiency of the algorithm. However, as the new links (and thus the new parameters on them) created by adding a new distance (i.e. adding 1 on k) may increase as k increases, which means the number of parameters need to be trained for the model may change on different k, this method may not be precise on some models. As a result, depending on the model and algorithms applied in the project, this evaluation method may need to be modified.

- The suggested value of k is optimal under different random training and validation sets, with the evaluation method developed in this project applied

  After suggesting an optimal value of k based on experiments, we should test that this conclusion holds under different settings of the model.

7. Ethical Considerations

I have read ethical guidelines listed in the module page of COMP390 [13] and will follow them strictly.

Data Sources & Protecting Personal Privacy

As mentioned before in Data Sources section, the original data of this project is retrieved from CSD [3], a database of public access, and then processed by the research group [4] before used as dataset. As the data is not related to any human individuals, and all algorithms that would be used to process data have been published, no particular confidentiality is needed during the usage of the data.

In addition, as no other human participants except supervisors, research members and the author myself are involved in this project, no considerations on the protection of personal privacy is required.

Model Development Based on Previous Research

As this project may take advantage of models, algorithms and techniques of previous research, all usage of these resources will be acknowledged and properly cited.

8. BCS Project Criteria

There are 6 outcomes of honours year projects required by the BCS, the Chartered Institute for IT [14]:

- An ability to apply practical and analytical skills gained during the degree programme.

  This project is based on a very specific and practical problem in crystallography. Practical and analytical skills are highly required in this project, as there is no standard answers for this project and the problem need to be solved in an original way.

- Innovation and/or creativity.

  Although there are similar research on this topic, the input data of those research is different from this project. This requires a new model to be proposed to solve the

problem.

- Synthesis of information, ideas and practices to provide a quality solution together with an evaluation of that solution.

  As stated in Key Literature & Background Reading, 3 categories of literatures will be covered; in Development & Implementation section, there is a primitive idea of the project development; in Testing & Evaluation section, 3 testing criteria were proposed to test the product model.

- Your project meets a real need in wider context.

  Prediction of element type based on its crystallographic environment is crucial to crystal structure validation and generation.

- An ability to self manage a significant piece of work.
- Critical self-evaluation of the process.

  As stated in the Project Plan and Risks & Contingency Plans section afterwards, strict plans and possible risks are proposed to manage the project, and to evaluate the process.

9. UI/UX Mockup

Since there is no user in this project, there is no possible user interface and user experiences.

10. Project Plan

There are in total 13 weeks for Design & Implementation phase of the project, in which 3 weeks are holidays and 3 weeks are exam weeks. Therefore there is approximately 9-12 weeks of time for the design and implementation of this project.

Since the specific model to use has not been decided and need further research, the project plan is considered under evolutionary model, which means the project development is based on the initial implementation, then continuously refine it according to the evaluation. Therefore modification & update of the model may happen during any stage in the design & implementation phase.

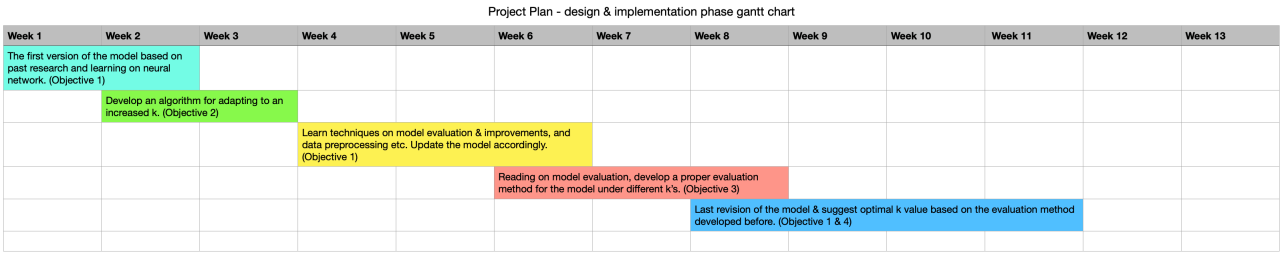
As mentioned in Aims & Objectives section, there are 4 objectives for the project:

- Objective 1: To develop a neural network model that is capable of predicting chemical elements given their k-NN distances.
- Objective 2: To develop an algorithm for a trained model to adapt to a larger k.

- Objective 3: To develop a method to evaluate the improvement of performance and the amplification of complexity of the model by increasing k.
- Objective 4: To suggest a best k value of complexity-accuracy tradeoff for the designed model.

Based on these 4 objectives above, 5 general steps are proposed to describe the model development process:

- T1: Reading & Learning on Neural Networks & Past research, and the first version of the model(1-2 weeks) (Objective 1)
- T2: Develop an algorithm for adapting to an increased k. (1 week) (Objective 2)
- T3: Learn techniques on improving a model, evaluating a model or preprocessing the data. Update the model accordingly. (2-3 weeks) (Objective 1)
- T4: Reading on Model Evaluation, develop a proper evaluation method for the model under different k's. (1-2 weeks) (Objective 3)
- T5: Last revision on the model & suggest optimal k. (2-3 weeks) (Objective 1 & 4)



Dependencies in the Gantt chart

Task ID	Description	Dependencies
T1	Based on research, produce a first version of the model	None
T2	Develop an algorithm to adapt to an increased k	T1
T3	Learn relevant knowledge and techniques, improve the model	T1
T4	Develop a proper evaluation method for the model under different k's	T1
T5	Last revision of the model, and suggest the optimal value of k.	T3, T4

Extra time is reserved in the design & implementation phase to handle possible risks. More details on how these steps would possibly be done have been explained in Key Literature & Background Reading and Development & Implementation sections.

Then, there are 3 weeks for testing & evaluation phase, during which the task is to test the final model based on the criteria proposed. Finally 3 weeks are reserved to prepare for dissertation and presentation.

11. Risks & Contingency Plans

Risks	Contingencies	Likelihood	Impact
The model is unable to gain acceptable performance on prediction.	Consult the supervisor and group members for possible reasons. Try reproduce other previous research's results. Try other possible models.	Low	High, the project may not be able to produce a reliable neural network model.
Time is not enough for learning adequate knowledge on chemistry and crystallography to design a comprehensible model.	Regard the model as a black box, and make modifications only based on modern deep learning techniques and past research.	Medium	Relatively low, as the model can have good performance even the network is not comprehensible.
I may encounter technical problems when using University's devices (failure to deploy development environment, or connection failure etc.)	Consult the supervisor and group members for technical assistance. Consult the IT services of the Department of Computer Science.	Relatively low	High, models will not be able to get trained, thus future improvements, testings and evaluations are impossible.
The project objectives and plans may not be adequate and/or accurate. There may be delays in development stages, and unexpected work may exist in implementation stage.	Maintain regular communications with the supervisor and group members, and do more literature review, especially at the beginning stage of the project, such that project objectives/plans could be revised in time.	Medium	Medium, the project may end up incomplete.

References List

[1] Widdowson, D. et al. (2021) ‘The asymptotic behaviour and a near linear time algorithm for isometry invariants of periodic sets’. Available at: https://arxiv.org/pdf/2009.02488.pdf (Accessed 23 Oct, 2021).

[2] Zhang, A. et al. (2021) ‘Dive into Deep Learning’. Available at: https://d2l.ai/index.html (Accessed 21 Oct 2021).

[3] Cambridge Crystallographic Data Centre, CCDC (2021) *Cambridge Structural Database*. Available at: https://www.ccdc.cam.ac.uk/support-and-resources/downloads/ (Accessed 23 Oct 2021).

[4] Kurlin, V. (2021) *Dr Vitaliy Kurlin: mathematics & computer science*. Available at: http://kurlin.org/index.php#group (Accessed 24 Oct 2021)

[5] Perkins, D. et al. (2020) ‘Crystallography’, in *Mineralogy*. Available at: https://opengeology.org/Mineralogy/11-crystallography/ (Accessed 23 Oct 2021)

[6] Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) ‘Learning representations by back-propagating errors’, *Nature*, 323(6088): 533-536.

[7] Deng, L. and Yu, D. (2014) ‘Deep Learning: Methods and Applications’, *Signal Processing*, Vol. 7, Issue 3-4, ISSN: 1932-8436. Available at: https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/?from=http%3A%2F%2Fresearch.microsoft.com%2Fpubs%2F209355%2Fdeeplearning-nowpublishing-vol7-sig-039.pdf (Accessed 23 Oct 2021)

[8] Schmidhuber J. (2015) ‘Deep Learning in Neural Networks: An Overview’, *Neural Networks*, 61: 85-117. Available at: https://arxiv.org/pdf/1404.7828.pdf (Accessed 22 Oct 2021)

[9] Ryan, K., Lengyel, J. and Shatruk, M. (2018) ‘Crystal Structure Prediction via Deep Learning’, *Journal of the American Chemistry Society*, 140(32): 10158-10168.

[10] Kittel, C. and McEuen, P. (2018) *Introduction to Solid State Physics*, 9th edn. New Delhi: Wiley.

[11] Pauling, L. (1960) *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*, 3rd edn. New York: Cornell University Press.

[12] Cordero, B. et al. (2008) ‘Covalent radii revisited’, *Dalton Transactions*, Issue 32, 2832-2838.

[13] Department of Computer Science, University of Liverpool (2021) *Ethical Conduct*. Available at: https://student.csc.liv.ac.uk/internal/modules/comp390/2021-22/ethics.php (Accessed 24 Oct 2021)

[14] BCS (2021) *BCS, The Chartered Institute for IT*. Available at: https://www.bcs.org (Accessed 23 Oct 2021)