

Metrics for Materials Discovery

With applications in the search for solid state electrolytes

by

Cameron Hargreaves

Supervised by Matthew Dyer & Vitaliy Kurlin



Preface

The list of people I would like to extend my personal gratitude to extends far beyond the character limit imposed by this thesis. It is the many brief encounters with great minds that have truly shaped the journey of this work, and I feel incredibly fortunate being able to experience this small slice of history that has allowed such rich connections of human experience to be shared across real and virtual spaces.

From Calderstones I would like to first thank John Cain and Darryl Ellson for inspiring me to take up my first non-musical passions, tolerating antics, and for rewarding my curiosity. In particular, I wish to thank Allan McLaughlin for the incredible impact you had on my life, I've thought of your words often.

Outside of academia I would like to thank Lucy, Kate, and Zoe for always looking out for me, your friendship eased this whole process. I needed to go for a lot of those walks. I will always remember the companionship Viilas, Tom, and Tom gave me, and I look back on our shared meals fondly. I would like to thank Andrew, Louis, and Joash for our invigorating times shared at sea. I've been incredibly lucky to be graced with the best sister Robyn, who has shown complete dedication to the success of all our projects. The weekly calls with Chris and Jill kept some semblance of normality through lockdown, and I am very thankful for the support the two of you and Roger have showered me with. Above all else, I would like to thank my amazing and creative parents Isla and Phil, for teaching me to walk my own path, to show kindness to others always, and to listen to the music.

At the University of Liverpool I would like to extend special thanks to my advisors John Claridge and Vladimir Gusev for the kindness and wisdom they have demonstrated throughout my PhD. Sam and Matt have been excellent people to co-brain with, and I've loved being able to bounce ideas off each other. I would like to thank Vitaliy Kurlin for his dedication to the training program that has enriched my life. Each of the co-authors listed before each chapter gave an incredible show of force coming together to assist each project, and I remain impressed at the collaborative efforts that are made possible at the materials innovation factory. It has been Professor Matt Rosseinsky's commitment in assembling such a high calibre team, along with his steadfast support whilst writing Chapters 2 and 3 that has driven many of my projects. Special notice should be paid to the efforts of Luke Daniels and Michael Gaultois for the unwavering patience they undertook when verifying work and teaching me how to write a scientific paper. Finally, I would like to thank Matthew Dyer for his never ending guidance and his presence when nothing could go right. Being an unceasing font of wisdom, Matthew has been the perfect mentor, and his exceptional judgement has seen each of these projects kept on track to completion, for which I will be forever grateful.

Abstract

Metrics for Materials Discovery: With applications in the search for solid state electrolytes by Cameron Hargreaves

The vast corpus of experimental solid state data has enabled a variety of statistical methods to be applied in high throughput materials discovery. There are many techniques for representing a material into a numeric vector, and many investigations apply the Euclidean distance between these vectors to judge similarity. This thesis investigates applications of non-Euclidean metrics, in particular optimal transport measures, or the Earth Mover's Distance (EMD), to quantify the similarity between two materials for use in computational workflows, with a focus on solid state electrolytes (SSEs).

Chapter 1 introduces the field of lithium conducting SSEs for use in batteries, as well as an introductory precursor for some of the machine learning concepts, for those without exposure to this field. The EMD is a function which returns the minimal quantity of work that is required to transform one distribution into another, and a tutorial on how to compute the EMD using the simplest known technique is provided given its relevance to later chapters.

In chapter 2 the discussion around the EMD is continued, and we introduce the workflow that has been developed for quantifying the chemical similarity of materials with the Element Movers Distance (ElMD). Given the affect that minor dopants can have on physical properties, it is imperative that we use techniques that capture nuanced differences in stoichiometry between materials. The relationships between the binary compounds of the ICSD are shown to be well captured using this metric. Larger scale maps of materials space are generated, and used to explore some of the known SSE chemistries.

At the beginning of the PhD, there were no substantial datasets of lithium SSEs available, as such chapter 3 outlines the lengthy process of gathering this data. This resulted in the Liverpool ionics dataset, containing 820 entries, with 403 unique compositions having conductivities measured at room temperature. The performance of leading composition based property prediction models against this dataset is rigorously assessed. The resultant classification model gives a strong enough

improvement over human guesswork that it may be used for screening in future studies.

At present, materials datasets are disparate and scattered. Using the ElMD in chapter 4, we investigate how different metric indexing methods may be used to partition gathered datasets of compositions. This enables very fast nearest neighbour queries allowing the automated retrieval of similar compounds across millions of records in milliseconds.

Chapter 5 introduces the technique Percifter for characterizing crystal structures, based on the principles of persistent homology (PH). This increasingly popular technique is used in materials science to describe the topology of a crystal. Percifter seeks to improve the stability of these representations for different choices of unit cells. These similarities may be observed directly, or compared through the EMD.

Table of Contents

| | |
|---|----|
| Abstract..... | 3 |
| 1. Introduction..... | 7 |
| Solid State Ionic Conductors..... | 8 |
| Machine Learning Prerequisites..... | 15 |
| Metrics..... | 24 |
| Optimal Transport..... | 29 |
| References..... | 37 |
| 2. The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions..... | 40 |
| Introduction..... | 39 |
| The Earth Mover's Distance..... | 45 |
| Pairing Structures to Compositions..... | 47 |
| Mapping Compositional Space..... | 50 |
| Binary Compositions..... | 52 |
| The Inorganic Crystal Structure Database..... | 59 |
| Principle Component Analysis..... | 61 |
| The Pearsons Database..... | 66 |

| | |
|---|-----|
| Conclusions..... | 82 |
| References..... | 85 |
| Supplementary Information..... | 88 |
| 3. A Database of Experimentally Measured Lithium | |
| Solid Electrolyte Conductivities Evaluated with Machine Learning..... | 120 |
| Introduction..... | 121 |
| Results and Discussion..... | 125 |
| Database Construction..... | 125 |
| Database Overview | 128 |
| Machine Learning..... | 134 |
| Supervised Learning..... | 143 |
| The Final Models..... | 157 |
| Conclusions..... | 161 |
| References..... | 164 |
| 4. Organising Materials Datasets for Similarity | |
| Queries with Metric Indexing Data Structures..... | 169 |
| Introduction..... | 169 |
| Metric Spaces..... | 172 |
| Metric Trees..... | 173 |
| Priority Queues..... | 176 |
| The MTree..... | 176 |
| The List of Clusters..... | 182 |
| Methods..... | 184 |
| Data..... | 184 |
| Tree Hyperparameter Selection..... | 186 |
| Results and Discussion..... | 187 |
| Indexing Times..... | 187 |
| <i>k</i> -NN Search Times..... | 188 |
| Range Query Search Times..... | 190 |
| Hosted Search Interface..... | 193 |
| Stress Testing..... | 196 |
| Conclusions..... | 198 |

| | |
|---|-----|
| References..... | 199 |
| Appendix 1 -Search Trees..... | 200 |
| 5. Percifter: A topological descriptor for periodic crystal structures..... | 205 |
| Introduction..... | 205 |
| Prerequisites..... | 208 |
| Betti Numbers..... | 209 |
| Homology..... | 211 |
| Vietoris-Rips Complex..... | 213 |
| Boundary Matrices..... | 213 |
| Computing Betti Numbers..... | 216 |
| Persistent Homology of Finite Spaces..... | 220 |
| Primitive Bravais Lattice Classification..... | 226 |
| Persistent Homology Shortfalls for Periodic Pointsets..... | 231 |
| Percifter..... | 233 |
| Results and Discussion..... | 238 |
| Empirical Construction Time..... | 238 |
| CPDs of Known Crystal System Relationships..... | 241 |
| The Ruddlesden-Popper Phases..... | 241 |
| The UiO MOFs..... | 244 |
| The Binary Compounds of the ICSD..... | 246 |
| Embedding Binary CPDs w.r.t the EMD..... | 250 |
| Topological Analysis of Solid State Electrolytes..... | 256 |
| Percifter EMD Distances..... | 264 |
| Rattled NaCl..... | 264 |
| Solid State Electrolytes..... | 267 |
| Embeddings of the Solid State Electrolytes..... | 271 |
| Further Work..... | 275 |
| Conclusions..... | 276 |
| References..... | 277 |
| 6. Conclusions and Further Work..... | 280 |

| | |
|---|-----|
| The Element Movers Distance..... | 280 |
| ML-Driven Design of Ionic Conductors..... | 284 |
| Percifter..... | 287 |
| Closing Remarks..... | 291 |
| References..... | 292 |

Introduction

Scientific endeavours involve interrogating a past corpus of knowledge. The tools we choose to carry out our investigations vary by field and practitioner, but the mental model is reliant on drawing similarities between past experiences.

“These similar elements could be substituted.”

“This dilemma is similar to a formal optimisation problem.”

“I subconsciously think rock salt and perovskite are more similar to one another than graphene, so I will not think about graphene.”

We don't typically note why our mental models form as we enact our curiosities. Gut feeling may be attributed to a great many success stories, and should not be ruled out as a valid technique for radical reinvention. But, survivorship bias is real.

The vast quantities of data that our investigations may now draw upon, means we must integrate the full computational capabilities made available to us. To guide the hunt in a more systematic manner, great strides have been made improving the accuracy of neural network models to predict materials properties, allowing pre-screening of many candidates prior to synthesis ¹.

Unfortunately, the inherently black box nature of the neural network means these predictions offer limited interpretability. Many have worked on improving this shortcoming ^{2,3}, but we can never truly understand our neural networks, as we are intellectually incapable of following the logical steps they take to formulate each of their outputs ⁴.

Capturing this knowledge, this *why* of a trends existence, remains one of the driving forces of discovery. A model may give us an answer, but we must apply our training if we wish to understand the mechanisms which rule our world. The field of data driven materials discovery has matured, and statistical methods now have

a firm place in the arsenal of the cheminformatician⁵. In this thesis we will present novel metrics to distinguish compounds in computationally guided materials discovery and demonstrate how these may complement machine learning (ML) algorithms in a modern workflow.

Given the breadth of potential topics the field of materials science encompasses, we shall tend to focus future discussions towards the study of solid state electrolyte (SSE) lithium ion conductors, suitable for use in the next generation of batteries. At the start of this investigation, whilst previous studies had been carried forward⁶ limited datasets necessitated the gathering of more data from literature sources. Over the course of this study, novel metrics were developed to interrogate and uncover trends within the Liverpool ionics dataset gathered, and we believe these metrics have utility beyond the investigations presented in this thesis. An introductory overview of some of the prerequisite topics is first provided.

Solid State Ionic Conductors

Energy demands of the 21st century compel us to improve the technology we use to store this resource. Batteries have enabled the mobile revolution, and current trade-offs between energy density, capacity, and cost have been highlighted as obstacles in the mass deployment of electric vehicles⁷ and renewable energy sources⁸. The lithium-ion battery has attained market dominance, but current designs are reliant on liquid electrolytes to transport Li^+ ions from the anode to the cathode during the discharge cycle⁹ (Figure 1). This liquid electrolyte is highly flammable, which means battery manufacturers must work within manufacturing tolerances to ensure there will be clear separation of electrodes over the full course of the cells lifetime, which reduces the overall energy density¹⁰. If the electrodes do make contact they form a short circuit, the battery will cease to function as the voltage will drop to

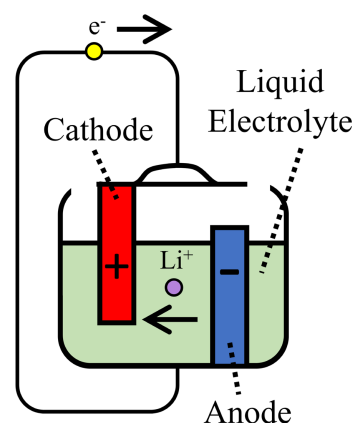


Figure 1: The mechanical principles of a liquid electrolyte battery cell, demonstrating the Li^+ ions travelling through the liquid electrolyte from anode to cathode during the discharge cycle.

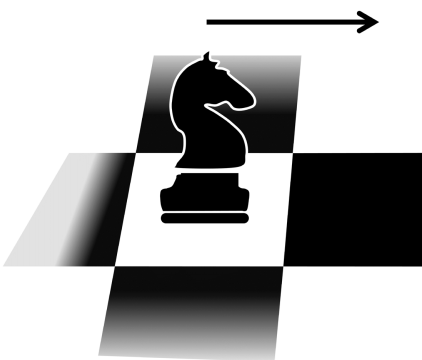


Figure 2: A magnetic chess piece snaps to a fixed position and requires energy to displace

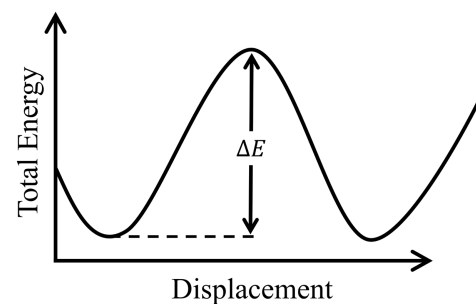


Figure 3: The total energy required to displace an ion from one energy state, i.e. position, to another in a crystal lattice.

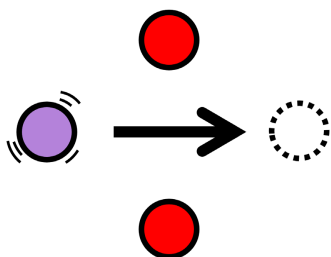


Figure 4: As a free ion gains kinetic energy through thermal motion, it may be able to push past the energy barrier and hop to a new site.

0, and the rush of current may set the liquid electrolyte on fire. By contrast, solid electrolytes are dense materials, which are inflammable, allowing for condensed designs. These also have the potential to utilise lithium metal anodes, which have a greater energy density than the lithium intercalated graphite anodes that the current designs use.

As heated liquid reagents cool into crystalline solids, the atoms of the crystal will settle into their most energetically favourable positions. The attractive and repulsive nuclear forces that bind structures together, will nudge each of the anions and cations into their lowest energy states forming a regular periodic structure. Only ions which are small enough to navigate their way through this backbone demonstrate mobile transport, such as Li^+ . At absolute zero temperature these ions will either take fixed positions in the repeating lattice, or be located more randomly at the relatively low energy interstitial sites found in between the periodic positions of the larger atoms of the structure.

When disturbed, a magnetic chess piece will stay in place (Figure 2). To slide it across a magnetic chessboard we must exert some force to move the piece away from the centre of each square before it snaps to the next. Similarly an ion in its lowest energy state requires energy to be added into the system to move the ion through a higher energy position before it loses this as it returns into an energetically favourable position elsewhere in the crystal. We call the amount of energy required to force an ion over the energy barrier into its next position the activation energy of the compound, ΔE , which is a fundamental property of the material (Figure 3).

If we heat the system above absolute zero, thermal motion will start providing enough energy to allow some mobile ions to jump this energy barrier. These ions want to be in low energy arrangements, and must carry enough momentum to force themselves through higher energy regions by thermally vibrating

faster and faster until jolted into a neighbouring low energy site (Figure 4). If an ion is not provided with enough energy to pass through an energy barrier, it will recoil and pass along some of its thermal energy to the wider structure. This motion is inherently stochastic, and when considered across the entire material the combined momenta of these ions will cancel, leading to zero net migration.

In SSEs, this mobile phenomenon can be regulated by applying an electromotive force (a voltage) across the solid material to induce a directed flow of ions. Several mechanisms have been identified which allow Li^+ ions to intercalate through the solid structure of an SSE when an electric field is applied across it, with interstitial hopping being the most common. Ions will flow in the direction of the electric field, which is from anode to cathode during the discharge cycle (Figure 5), and in the reverse direction whilst charging. The most common method of inducing this voltage is through the use of an electrochemical cell.

In an equilibrium solution, if Li atoms encounter Co^{4+} , the Li is likely to donate its electron to produce the more energetically favourable Li^+ and Co^{3+} . By stuffing certain cobalt containing compounds with extra Li atoms we may expect this reaction to take place inside the crystal, which is what happens in the cathode of commercial battery cells which use LiCoO_2 cathodes (Figure 6). The favourable Co^{3+} oxidation state means the cathode material will attempt to draw in Li^+ ions through the electrolyte during discharge, inducing a potential voltage. In parallel, an oxidation of Li to Li^+ will take place at the anode to produce the ions which intercalate through the electrolyte (Figure 7). However, the system will be unable to, as the additional force applied by the excess of positively charged particles in the cathode will not allow for further electron transfer between nuclei to take place.

Excess negative charge in the anode will not allow any Li^+ ions to leave, and the reaction stops and remains in a stable condition. This

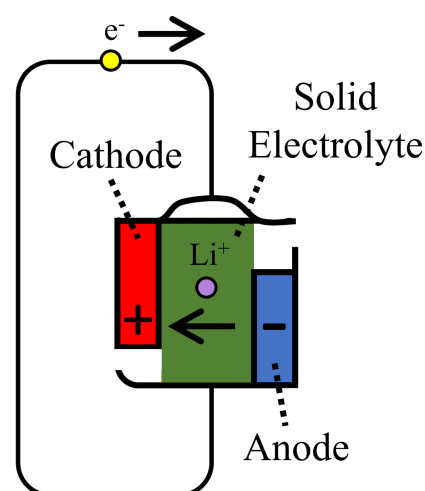


Figure 5: The mechanical principles of a SSE battery cell, demonstrating the Li ions intercalating through the solid electrolyte from anode to cathode during the discharge cycle.

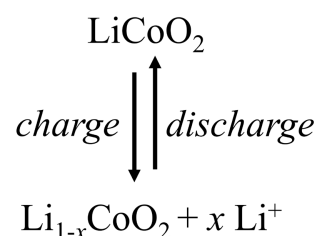


Figure 6: The charge (down arrows) and discharge (up arrows) reactions that take place at the cathode in a LiCoO_2 cathode Li anode SSE battery.

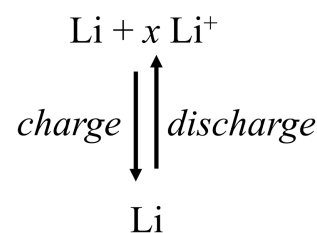


Figure 7: The charge (down arrows) and discharge (up arrows) reactions that take place at the anode in a LiCoO_2 cathode Li anode SSE battery.

is the state of a charged battery cell when in storage. Connecting a wire between the two electrodes allows electrons to flow from anode to cathode and recombine with the Li^+ ions as they're deposited. This reaction then continues producing usable electricity in the discharge cycle, until all Li atoms in the anode have been depleted, the cathode can accept no more Li, a mechanical failure, or the circuit is disconnected. When the battery has fully discharged, the cell must be recharged by applying an external electric field, which forces the Li^+ ions to go in the reverse direction.

One of the difficulties that arise in this setup is choosing a suitable material for the electrolyte ^{11,12}. It must be chemically stable, otherwise it may begin forming contaminants with the anode. It should be hard, otherwise lithium deposits may build up and puncture the electrolyte. It should be malleable to absorb mechanical stresses. Ideally it would be stable in air, as this makes manufacture much simpler, and certain elements are more favourable than others in regards to cost, environmental impact, and concerns over ethical sourcing.

Whilst glassy and polymer type materials may be used as the transport medium in Li SSEs, the high entropy disorder present in these materials makes these more difficult to characterize structurally. As such, we shall focus our discussion on ceramic type materials which form periodic crystalline structures of anionic backbones, often containing interstitial sites of partial occupancy. There are three broad compositional families these materials commonly fall into, oxides, sulphides, and phosphates, each with identified strengths and shortcomings.

Oxides cover the garnet type structures with the first discovered being $\text{Li}_5\text{La}_3\text{Ta}_2\text{O}_{12}$ ¹³ (Figure 8). Garnets are chemically stable when placed against anode materials, and the relatively low conductivity of $3.4 \times 10^{-6} \text{ S cm}^{-1}$ ¹⁴ was improved by altering the composition to $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$ (LLZO) where a conductivity of 4

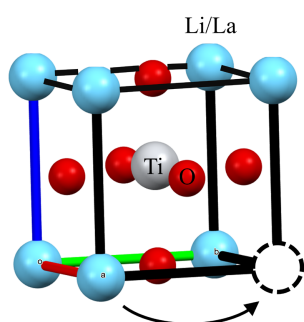


Figure 8: LLTO perovskite structure with a structural defect allowing ion hopping

$\times 10^{-4} \text{ S cm}^{-1}$ may be attained at room temperature (r.t.)¹⁵. This is not as high as the perovskite material $\text{Li}_{0.3}\text{La}_{0.53}\text{TiO}_3$ (LLTO) with a conductivity of up to $4 \times 10^{-3} \text{ S cm}^{-1}$ ¹⁶. LLTO is additionally stable against the anode, but unfortunately the grains that these crystals form introduce a large resistance as ions cross the grain boundaries. This significantly reduces the total conductivity, as low as $1.7 \times 10^{-6} \text{ S cm}^{-1}$ ¹⁷.

Li Super Ionic CONducters (LISICONs¹⁸) based on Li_4XO_4 and Li_3XO_4 ($X = \text{Al, S, Si, Ge, Ti, P}$) exhibit a fairly low conductivity at room temperature, and whilst these do remain stable in humid air, which allows for easy manufacture, these are not stable against lithium anodes. The lowest impedance LISICON has so far been achieved by adding a zinc dopant and a germanium cation to the formulation giving $\text{Li}_{2.8}\text{Zn}_{0.6}\text{GeO}_4$. A popular material that has been assessed to minimise interfacial anode interactions is based on the LISICON phosphate material Li_3PO_4 ¹⁹, processed through magnetron sputtering in the presence of N_2 gas to create $\text{Li}_{2.9}\text{PO}_{3.3}\text{N}_{0.46}$ (LiPON²⁰). Whilst this compound has a fairly poor conductivity of $1.4 \times 10^{-6} \text{ S cm}^{-1}$, it has high stability against lithium metal and is easy to manufacture making it suitable for use as a barrier layer to protect the surface of the anode.

Other phosphates such as the NASICON (Na Super Ionic CONductor) were some of the earliest investigations into mobile ion transfer in the solid state, based on the $\text{Na}_{1+x}\text{Zr}_2\text{Si}_x\text{P}_{3-x}\text{O}_{12}$ system²¹. Substituting Li and exploring other dopants means the associated family of Li electrolytes is well explored, and principally characterised by $\text{Li}_{1.3}\text{Al}_{0.3}\text{Ti}_{1.7}(\text{PO}_4)_3$ (LATP²²), which has a relatively high r.t. conductivity of $3.5 \times 10^{-4} \text{ S cm}^{-1}$. Unfortunately, this material is not stable against the anode, as the Ti will reduce the Li. The stability of this interface can be improved by substituting Ti for Ge in $\text{Li}_{1.3}\text{Al}_{0.3}\text{Ge}_{1.7}(\text{PO}_4)_3$ (LAGP²³), but this possesses a conductivity which is an order of magnitude lower than LATP.

Given the slightly lower electronegativity, sulphur forms weaker bonds with Li^+ ions than oxygen, allowing them to pass through the structure with less resistance. SSEs based on a sulphide anionic lattice, such as Li_3PS_4 , display significantly higher conductivity than oxides, competitive even with their liquid counterparts.

$\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ (LGPS²⁴) exhibits the highest room temperature conductivity of $1.2 \times 10^{-2} \text{ S cm}^{-1}$, with the $\text{Li}_6\text{PS}_5\text{X}$ ($\text{X} = \text{Cl, Br, I}$) family of argyrodite structures possessing a conductivity around $1 \times 10^{-3} \text{ S cm}^{-1}$ ²⁵. These are very soft materials, which can be a favourable property as it allows the battery cell to accommodate the changes of volume that take place at the electrodes as they absorb and secrete lithium, reducing mechanical fatigue.

Unfortunately, this flexible nature is also to their detriment, as dendrites of lithium metal may form as lithium is deposited on the anode. Given the softness of the electrolyte material, these deposits can easily push their way through the electrolyte until the battery forms a short circuit, ceasing current flow (Figure 9). Further, sulphides are not stable materials in air nor against lithium metal, making them difficult to manufacture, and challenging to distribute.

The interest in discovering new materials which have sufficiently high conductivity, chemical stability against Li, and good mechanical strength means that many hundreds of candidate materials have been reported in the literature, with many more synthesized and discarded as uninteresting. Discovering whether a novel synthesis candidate has already been examined and reported can be a lengthy task, in no part due to the highly fragmented landscape of materials databases. This specific problem of cataloguing previously reported materials in a chemically meaningful way is what instigated the development of the EIMTree application as a means of indexing these records, which is discussed further in Chapter 3.

For SSE batteries there are commercial offerings available, but to date, no single electrolyte formulation has proven itself sufficiently

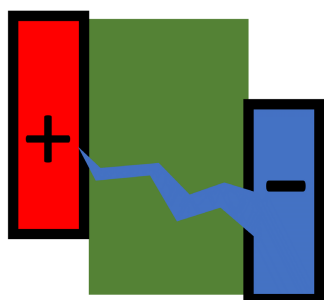


Figure 9: An illustration of the process of dendritic growth piercing the solid state electrolyte, short circuiting the battery.

superior to others that it has achieved market dominance, nor widespread acceptance as a replacement for liquid electrolyte. Thankfully, there remain plenty of avenues to dope and substitute elements within the known structure types to search for new materials with favourable properties. Given the additional complexity introduced when examining multiple properties simultaneously, and the importance of ionic conductivity to the perceived quality of a new electrolyte material, ionic conductivity will be the target property that we focus our discussion on, with more conductive materials deemed “better” materials.

The discovery of an entirely new structure type is a cause for celebration, given the flurry of activity that accompanies such a finding, searching for new substitution and doping pathways that can be applied. Theorising which substitutions could lead to stable materials requires training and subject matter expertise. Humans can interpolate their past experiences together to hypothesise experimental setups which lead to promising outcomes. However, due to the complexities of these systems, it is a very difficult task to predict which elements will interact in specific dopings to give increased performance in target properties.

The challenge of correlating properties across many different materials is twofold. Firstly, there is very little data available. Experiments take a long time to run and validate results, and a lack of interest in publishing unremarkable materials means that the literature will be restricted, and anthropogenically biased towards high performing materials. Secondly, as computational materials design remains a field in infancy, there are few established numeric protocols to process large quantities of materials science data and extract meaningful relationships. The core aim of this thesis will be examining computational and statistical methods which we can use to consider these spaces of materials, allowing us to draw chemically insightful observations from the data that is available.

Machine Learning Prerequisites

Glimpses of the human ability to notice and exploit patterns date back as far as 21,000 B.C. on the shores of the sea of Galilee²⁶. The earliest recognised signs of agriculture take the form of compost piles and wheat stones, enshrined by rising waters and revealed by drought. Our ability to recognise and exploit simple patterns, such as observing the seasons or the act of planting a potato to grow more potatoes, is one of the key traits that separates us from many other species. Furthermore, we possess the ability to abstract ideas into complex observations. If potatoes are planted during each of the four seasons these plants can be compared to see which bear the most fruit. This information can be used to guide and refine future sowing schedules, allowing the health of the villages population to flourish. These historic luminaries laid the foundational stones of statistical learning through observation, patience, and spoken word tradition. Now we have computers.

The refinement of various techniques, algorithms, and procedures that today fall under the data science umbrella tells a several millennia story. To ensure the necessary material and terms have been introduced for later chapters we shall focus our discussion on the topic of unsupervised learning and embedding, as there is a natural partnership between these techniques and the materials based metrics introduced later. We shall briefly review supervised learning, as our metrics may be integrated with such methods. This shall also introduce the topic for Chapter 3, where we carry forward an investigation predicting the ionic conductivity of materials based on their compositional makeup. An accelerated executive summary is required to ensure that a broad exposure is provided to those with no prior exposure to ML. Those familiar with these topics may skip this introductory, and historically inaccurate, summary to the following section.

Let us tackle some of the problems our first agricultural ancestors may have faced using the language and methodology of the

modern day. The first piece of information which must be determined beyond reasonable doubt is whether a plant is edible, or whether it is poisonous. Being an industrious hunter gatherer, you have kept track of the leaves of all the plants you and your friends have eaten. You order these leaves by their width and height on the ground (Figure 10).

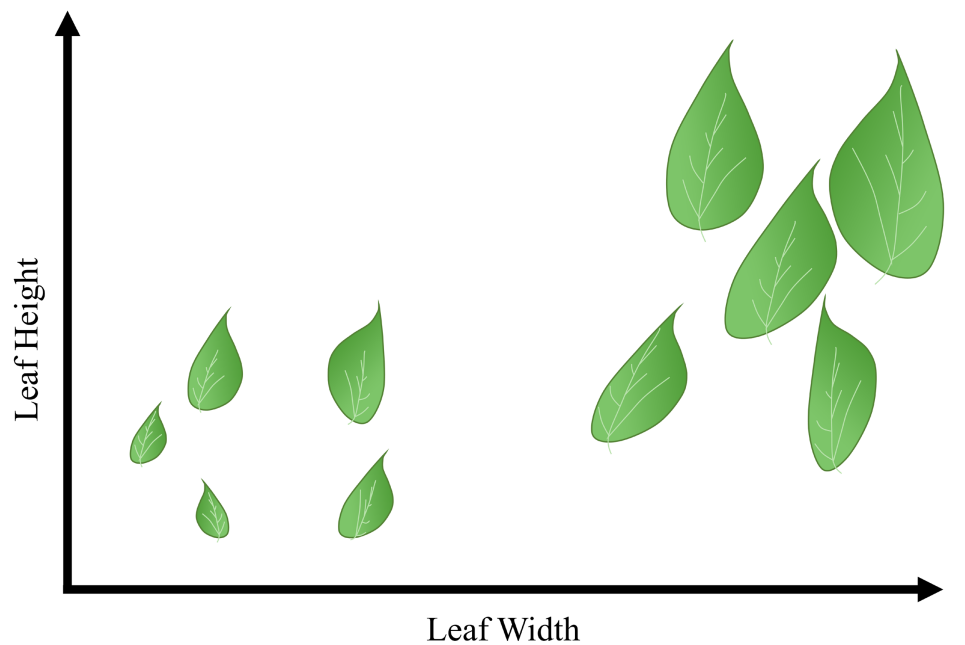


Figure 10: An empirically observed embedding of the properties of fauna in respect to their leaf sizes

With the support of your family members, you conduct an empirical study into the effects of consuming each plant, and make note of the plants which can be eaten only once (Figure 10).

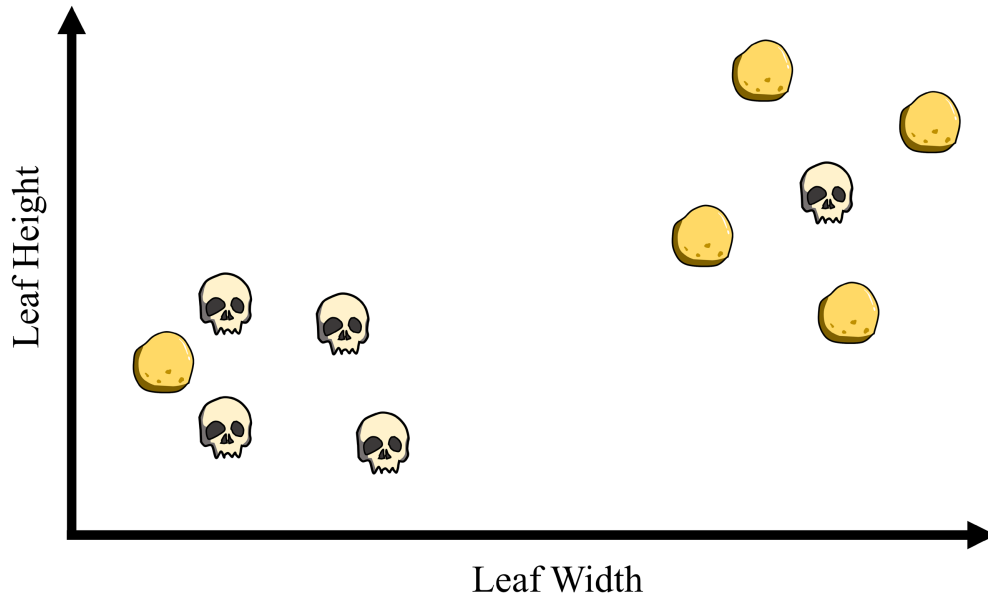


Figure 11: Each plant from Figure 9 arranged by leaf width and height, labelled with a potato icon if it is edible and a skull icon if it is poisonous.

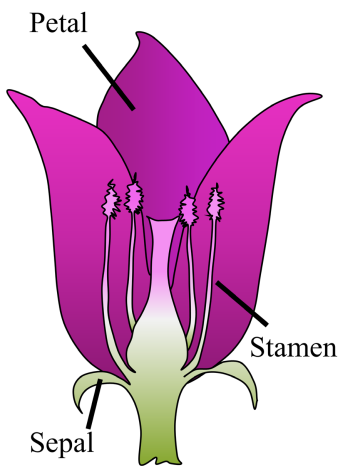


Figure 12: A diagram of the petal, sepal, and stamen locations of each plants flowers.

It becomes apparent which sort of plants we should focus future investigation on, but we do not want to poison our friends. You find seedlings for each of the vegetables, observe the plants through their life cycle, and gather the flowers they produce. There is now data for the petal width and height, sepal width and height, as well as the stamen height (Figure 12). Unfortunately this gives us seven numbers to represent each plant, and we may no longer order these on the ground to follow the pattern. We can use unsupervised machine learning to give us more valuable information about the data in our possession than each of these numbers do in isolation.

Embedding techniques focus on reducing the dimensionality of a dataset to two components (axes). One of the oldest and most widely used techniques is Principle Component Analysis (PCA). This uses techniques from linear algebra to identify potential axes of the dataset which follow greatest variance, applying linear

transformations to the data to squash the data into a lower dimensional space, using these axes of greatest variance as the new coordinate system. If many properties combine into a clearly identifiable trend, these trends should be captured when we change the co-ordinate system to the principle components. The data that we have gathered enables us to represent each plant by a vector of seven numbers. As trends in 7-dimensional space are difficult to visualise and identify, we shall reduce these vectors to 2-dimensional co-ordinates with PCA (Figure 13).

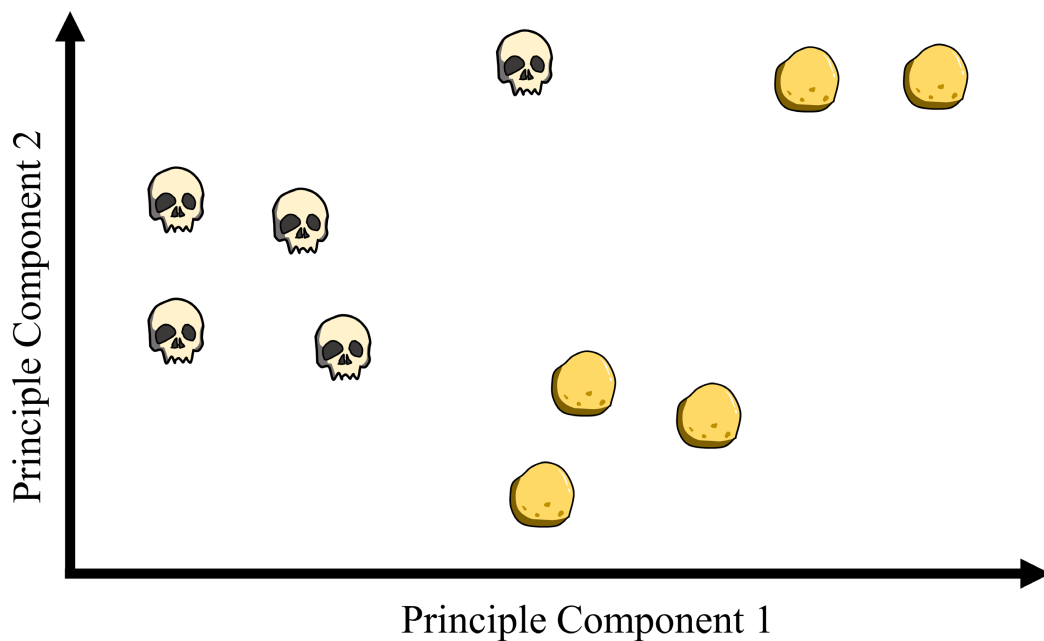


Figure 13: The same objects and labels given in Figure 11, with new positions obtained by projecting the seven pieces of information about each plant to two new co-ordinate axes. When this process is carried forward, we want the relative distances between each object in lower dimensional space to reflect the position each object has in relation to the other objects in high dimensional space.

With the embedded data we can see that there are distinguished characteristics which clearly separate the edible plants from the inedible. If we encounter a new plant on our travels, we may now apply the same linear transform to judge whether it is safe to eat before volunteering a sibling.

Using your new method of identifying edible food you decide to plant these five species in a neighbouring field. Sixty nights later you do this again, and again every two months for the remainder of the year. At the end of the plants harvest you take note of the volume of food each plant has produced. By plotting the month of seeding against this volume of food, trends may be identified. We may predict what volume of produce a plant is expected to produce when planted at a given time period by fitting a regression line to the data (Figure 14). Here we create a simple trend line for the data by taking the average value of each time period and overlaying a curve to these averages.

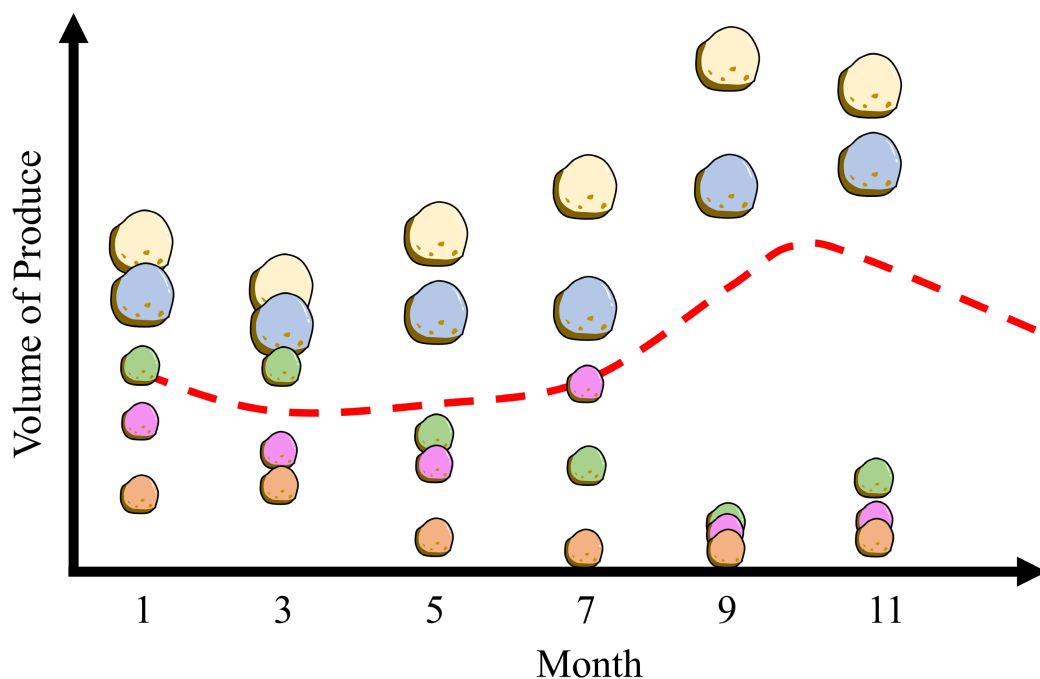


Figure 14: The edible plants and the produce each plant produces when seeded in a given month of the year. Average overlaid in red.

Looking at this trend and following the path to its maximum point, we can find the best time of the year to plant seeds. We hypothesise that if all seeds are planted in the tenth month of the year, the total volume of produce will be maximised. Executing this plan leads to greater food stocks the following year, but you believe this situation can be improved.

In simple examples such as this, we have the capacity to evaluate optimal planting strategies for each plant by eye. For larger domains when we are dealing with millions of data points this isn't a feasible strategy. Five lines could be drawn between the points of each unique species and used to judge the best time of the year to plant each plant (Figure 15).

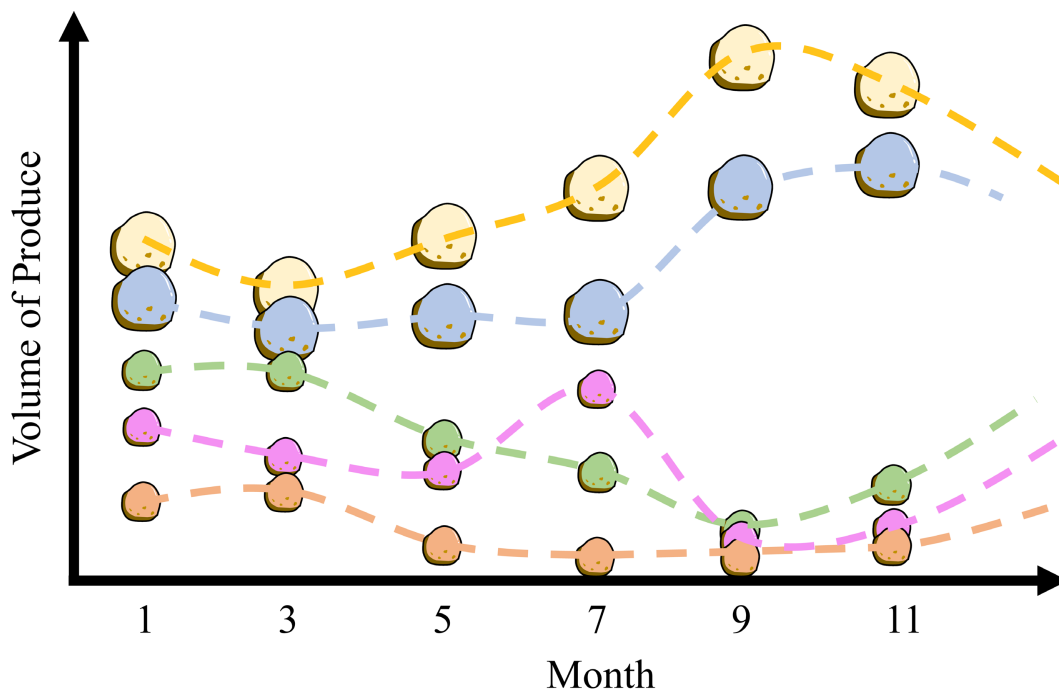


Figure 15: A multi-set of regression functions, each of which have been overfitted to the given data

With our limited seedlings, and thus datapoints, this approach is highly sensitive to outliers. Following the pink series we see that the seventh month would be recommended for planting. Despite this being what the statistics suggest, from the other data points we see why we might not believe this to be a sound judgement. We would say that the pink trace has overfitted to the data, as the regression line sticks too closely to the outlying points and does not follow the general trend.

A more robust solution would be to use the information that is provided to us by the plants which are very similar. We can see in

our embedding that there are two clusters of edible plants with shared physical characteristics. Let us use another unsupervised machine learning method to assign a numeric cluster label to each point in the embedding, which will be used to filter and consider each cluster separately. Automated clustering techniques allow us to gather together objects with shared features and assign them as being part of the same set. If we have no prior information about an object other than its high dimensional vector representation, this can be very useful as it tells us which previously reported objects we should begin referring to when characterising the new object. The clustering technique we will use in this thesis is called Density Based Spatial Clustering of Applications with Noise (DBSCAN ²⁷).

This algorithm operates by examining the regions of density in an embedding, and assigning a cluster label to each distinctly separable cluster, based on a cut-off radius. A coarse simplification of the approach may be constructed by overlaying a fixed radius covering disk centred at each point in the embedding (Figure 16).

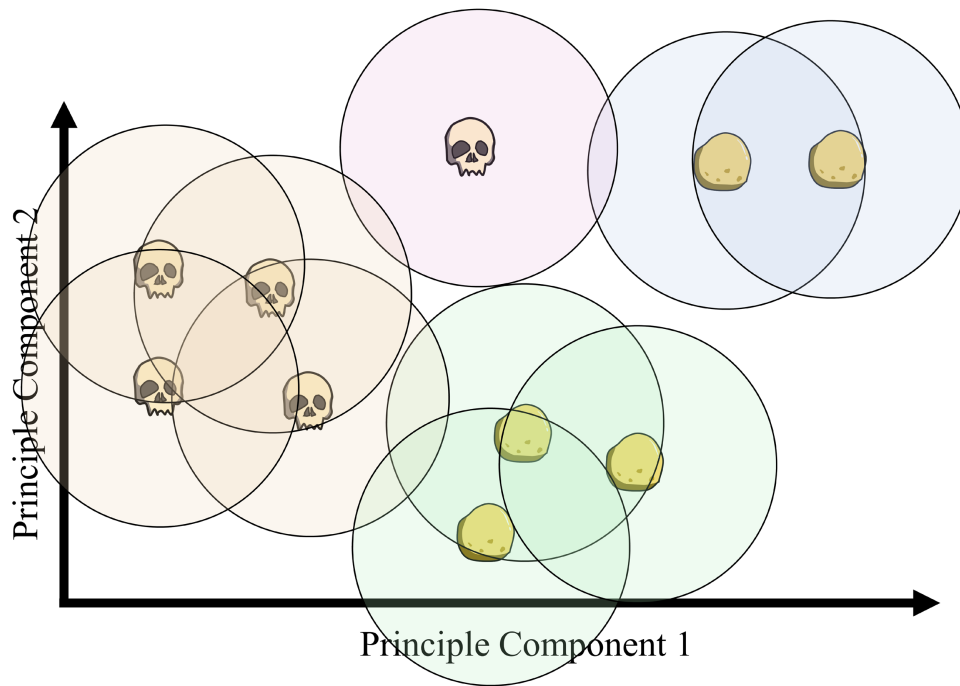


Figure 16: A simplified example of the operating principles of the DBSCAN algorithm. A disk is overlaid on each datapoint with a fixed size radius. If a point is contained within the disk of another point, then we say each of these points belong to the same cluster.

When doing this, for an appropriately sized choice of radius, several of the points covering disks will intersect other points. When this happens we shall assign these to be a part of the same cluster. As a consequence, contiguous regions of the space will all be assigned to be a part of the same cluster. This allows us to draw similarities between points which may not be in close proximity to one another on the embedding, but still possess logical connections through familiar neighbours and longer range trends. With cluster labels in hand, we may assign a regression curve to each of the two sets of edible plants. Depending on the plants classification, we arrive at different seasons for optimal planting. Following through with this plan leads to an increase in crop yield the following year.

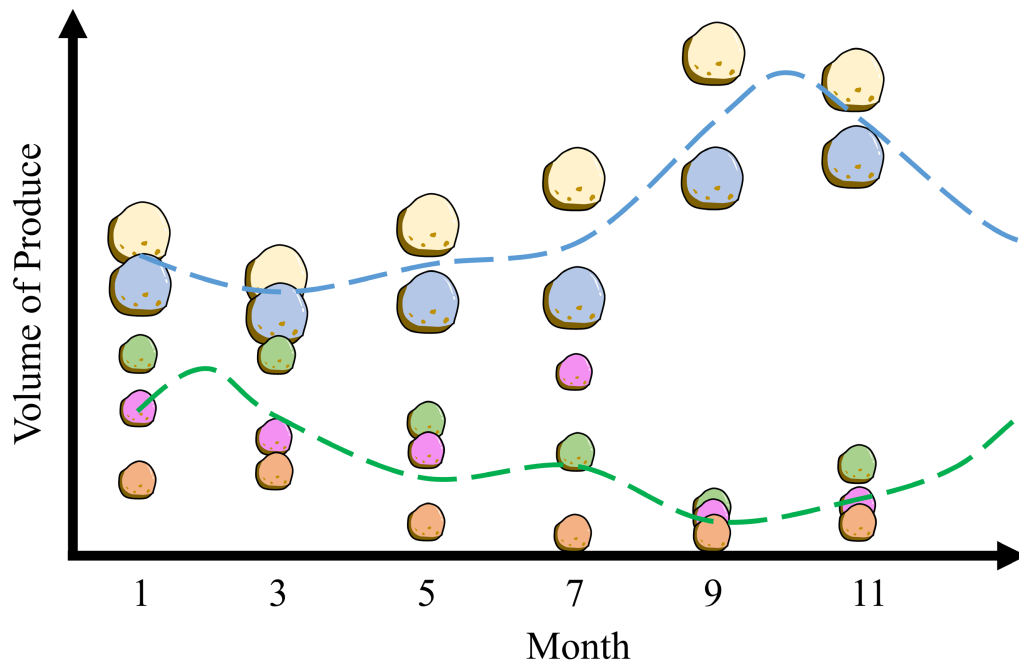


Figure 17: An appropriately chosen set of regression functions to predict the planting season for the dataset.

We have established the best method of growing each edible plant through the use of a regression model, but not discussed the inner workings of how this single variable model takes in the month and predicts the quantity of produce. The way each model architecture assigns a real valued number for an input of potentially many variables is dependent on the mathematical recipe which is followed by the algorithm. In our work we shall focus on the applications of neural networks and random forests.

Many regression models can often be converted into classification models through minor architectural modifications. Instead of returning a real valued number, these instead return an integer label predicting the set that an object is expected to belong to. These are normally supervised models, which means the model must be exposed to numeric representations of each object with their associated target classification labels that we will aim to predict. These models are beneficial as they need not be restrained to the

potentially lossy lower dimensional spaces which we have thus far constrained ourselves to for the sake of deductive reasoning. We expect that embedding a space to a lower dimensional representation will discard some of the information about the space, losing some of the more nuanced relationships present in its high dimensional structure. This could make forming predictions about each object a more difficult task for our models.

In this example, we have a seven dimensional vector for each plant, and four class labels which we are trying to assign each plant to (the target value). We may fit a classification model to this training data, and then when new plants are encountered in the wild we can measure their physical characteristics, predict whether they are edible with our classification model, and decide the best season in which to plant them with the appropriate regression model. This allows us to exercise a degree of caution when sampling new plants, and maximises the impact of the efforts exerted when planting new species.

Metrics

Assigning a class label to previously unseen vectors requires some method of measuring similarity to previously observed samples. A simple technique could be taking the k nearest neighbours in the training set to a query, and returning the most common class label. We must define a distance function, or metric, to ascribe these notions of “near” and “far”. There are many metrics which may be applied between two vectors, and the expressivity that each metric carries is often domain specific. For example, the Hamming distance is very useful for binary vector comparisons, but does not help distinguish vectors of real numbers. The most commonly used metric, and the one that is most familiar to us, is the Euclidean metric. Between two objects, O_x and O_y , in 3-dimensional space with the co-ordinates (x_1, x_2, x_3) , and (y_1, y_2, y_3) the Euclidean distance between these two is:

$$d(O_x, O_y) = \sqrt[2]{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

For higher dimensional spaces, such as the 7-dimensional space we have quantified our plants into, the sequence may be continued indefinitely.

$$d(O_x, O_y) = \sqrt[2]{(x_1 - y_1)^2 + \dots (x_7 - y_7)^2}$$

Under this metric, we say that our space observes Euclidean geometry, and each of our vectors may be considered as points in this high dimensional space. However, there are cases where the Euclidean metric may not be the most expressive metric.

Expressive is a subjective word, and in many domains the Euclidean distance *will* express similarity and dissimilarity. Under the Euclidean distance, if two vectors are identical they will always have a distance of 0, otherwise the distances will be greater than 0, and in 3-dimensional space this metric is perfectly valid and fully aligns with our interpretation of straight line distance.

Unfortunately, for higher dimensional spaces, such as the vector spaces we will use to represent our materials in later chapters, this interpretation begins to drift.

The geometry of higher dimensional spaces means that unless points fall very close to one another, as we increase the number of dimensions there will be an increasingly large distance between points, and points will become increasingly difficult to distinguish from one another. This is referred to as the curse of dimensionality, and many successful techniques in classical machine learning can ascribe a portion of their performance from successfully thwarting this phenomenon.

The Euclidean distance takes each of the axes in turn and considers the interpoint similarity along the axis. An identified shortcoming of this metric is it cannot take any inter-column similarities into account. For certain objects which are represented by vectors, such as the vectors of real numbers that represent the heights of bars in a discrete probability distribution, neighbouring columns may be

very similar to one another. Judging the similarities between column features will require some further mathematical machinery, but first some formalisms.

Thus far we have casually used the terms similar, distance, and metric, as these are common terms. For convenience, we shall set a simple bound for similar and then provide definitions for the other terms. If we take a query object and place all other objects in our possession in order of their similarity to the query, then the 15 closest objects shall be defined as the most similar to the query. Note this may not mean they *are* similar, as they could simply be less dissimilar than the other objects in the set, but throughout this work we shall use this as the cut-off for similar. This leaves the ambiguous question of what is similarity?

We need a numeric quantifier if we are to determine similarity between objects of large datasets, ideally a metric. We may assign a function, d , which will take in two objects and return a numeric quantifier to discern them, but to be called a metric this function must satisfy the metric axioms. Being a metric is of great benefit, as many statistical methods used in our analyses are generalisable to a range of tasks, but only if the similarity measure is a metric. Fortunately, these formalisms are brief, simple to follow, and generalise to any metric space. Let us use a simple space to demonstrate these axioms before moving onto the materials domain. The integer scale, represented by the symbol \mathbb{Z} , being all the whole numbers from minus infinity to infinity, shall be the first space we shall study to introduce the rules of metric spaces.

For each object, O_x with an associated value of $x \in \mathbb{Z}$, we want to know if a second object O_y with $y \in \mathbb{Z}$ is the same, similar, or far away. To do this we define a similarity function, $d(O_x, O_y)$, which shall return the distance between two objects in our space. For the integers, the metric which aligns with mathematical intuition takes the absolute magnitude between the two values:

$$d(O_x, O_y) = |x - y|$$

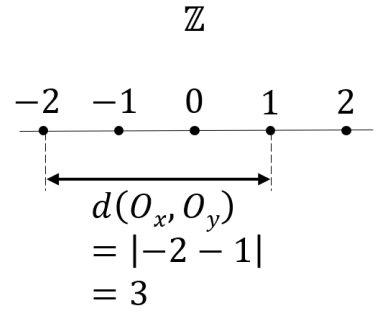


Figure 18: The integers contained in the set \mathbb{Z} on a number line, demonstrating how a distance function which takes the absolute difference between two numbers can be used as a metric of similarity.

$$d(O_x, O_y) = 0$$

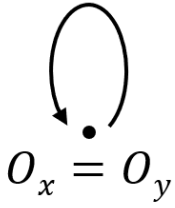
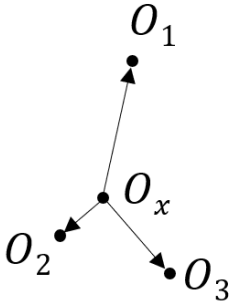


Figure 19: If the distance between two objects is zero the objects are equivalent under the chosen metric.



$$d(O_x, O_i) > 0$$

Figure 20: If the distance between two objects is greater than zero, the objects must be different under the chosen metric.

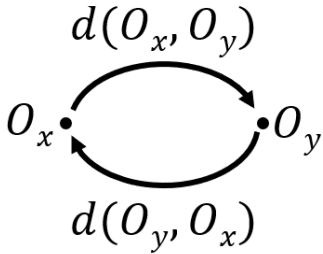


Figure 21: If the distance $d(O_x, O_y)$ gives one value, then $d(O_y, O_x)$ should return the same value.

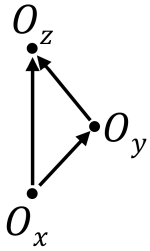


Figure 22: The triangle inequality states that the straight line distance $d(O_x, O_z)$ must be equal or smaller than the total distance of any path from O_x to O_z that detours via a third point O_y .

What values may this function take? If this returns a value of 0, then x must be equal to y , and we would say that these two objects appear the same under the metric. The first axiom of metric spaces (Figure 19) states that the distance between any object and itself must be equal to 0:

$$d(O_x, O_x) = 0$$

$$d(O_x, O_y) = 0 \leftrightarrow O_x = O_y$$

The next rule, the rule of positivity, states that the distance between two distinct objects must always return a positive value (Figure 20). It is clear that d satisfies this condition, as taking the absolute magnitude ensures the result will be non-negative. Many algorithms will make assumptions about distance and proximity which would fumble with negative distances, as these make do not make any physical sense.

$$d(O_x, O_y) > 0 \leftrightarrow O_x \neq O_y$$

The rule of symmetry ensures that the distance function does not change its output when we permute the order of the inputs (Figure 21). We must have consistency in distances when navigating between objects.

$$d(O_x, O_y) = d(O_y, O_x)$$

The final axiom formalises the notion that the shortest distance between two points is a straight line. This is true in 3-dimensional space, and also true in any metric space. If we take a straight line path, called a geodesic in metric spaces, between two objects O_x and O_z in the metric space, the total distance travelled must be extended if we take a detour via a third object, O_y , that does not fall on the original path (Figure 22). This is commonly referred to as the triangle inequality, as the edges between the points form a three edged shape.

$$d(O_x, O_z) \leq d(O_x, O_y) + d(O_y, O_z)$$

The triangle inequality must hold for any three objects in the metric space, and all four axioms must be satisfied for the distance function to attain metric status. Demonstrating why d is a metric on \mathbb{Z} , let us substitute the right hand side of the triangle inequality with the definition of d :

$$d(O_x, O_y) + d(O_y, O_z) = |x - y| + |y - z|$$

From the definition of magnitudes we know that $|x| + |y| \geq |x + y|$. We can substitute $(x-y)$ and $(y-z)$ into this property to give:

$$|x - y| + |y - z| \geq |(x - y) + (y - z)|$$

Through associativity we may cancel the $-y$ and the y variables giving

$$|x - y| + |y - z| \geq |x - z|$$

As $|x - z|$ is the definition of $d(O_x, O_z)$ we can see that the triangle inequality must hold for this function on all values of \mathbb{Z} .

Just because a distance function is a provable metric, does not mean that it expresses similarity with nuance or fidelity. The discrete metric is the canonical example which can be applied to any set of objects and is a valid, if somewhat limited, metric. This will simply assign a distance of 1 between all non-identical objects. The distance function for this metric is therefore:

$$d(O_x, O_y) = \begin{cases} 1, & x \neq y \\ 0, & x = y \end{cases}$$

This certainly passes the first three axioms, and indeed passes the triangle inequality, as $d(O_x, O_y) + d(O_y, O_z)$ will be equal to 2 when O_x, O_y , and O_z are distinct objects. It is clear the discrete metric is not an expressive metric however. This distinguishes objects of the space, but disregards a majority of the information that has been provided to us by the object. None of these distances will vary in magnitude from one another, $d(2, 3) = d(2, 10) = d(2, 10,000) = 1$ (Figure 24). We tend to have firm opinions of the magnitudes that distances between the integers ought to take, as we

$$\begin{aligned} |x| + |y| &\geq |x + y| \\ |-3| + |6| &\geq |-3 + 6| \\ 9 &\geq 3 \end{aligned}$$

Figure 23: A simple demonstration that the property $|x| + |y| \geq |x + y|$ is true when $x=-3$ and $y=6$.

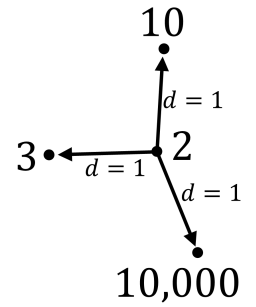


Figure 24: The discrete metric does not distinguish dissimilar objects with nuance.

call this counting. This highlights that just because a measure is a metric, it may not be an expressive metric.

It is easy to spot when we have chosen the “wrong” metric in these simple and exaggerated cases, but in more complex settings established metrics may attain acceptable distinguishing performance for similar objects, but fail to distinguish dissimilar objects coherently from one another. For us to extrapolate relationships across chemical domains, we need to be able to recognise local patterns that are happening in nearby regions of chemical space as well as global patterns which could be occurring in dissimilar regions of chemical space. If the underlying metric being used to differentiate points in a space is incapable of recognising long range patterns because all dissimilar points look the same, we have not chosen an expressive metric.

One of the core enquiries presented in this thesis will be exploring how certain representations of materials may benefit from abstracting the problem into the domain of mass distributions and mappings between distributions. Instead of examining the absolute magnitude between each variable, as with the Euclidean distance, the minimal cost to transport mass from one distribution into a second is taken, ascribing the total quantity of work required as the “distance” between the two distributions. This requires a more computationally intensive algorithm than the Euclidean distance, but can highlight similarities the Euclidean distance misses.

Optimal Transport

Optimal transport measures provide a consistent numeric evaluations of similarity between mass distributions, which are sensitive to small variations for both similar and dissimilar distributions. It has seen applications in other areas of materials science between SOAP²⁸ and AMD²⁹ structural descriptors. In this thesis we shall discuss some novel approaches where this technique may be applied.

The problem of finding the best method of transporting mass, is one where humans possess strong instinctive reasoning. If you have dug out a potato trench and you have excavated the earth in a row next to your trench, then you know that it is better practice to fill in the trench with the pile of earth that is closest to each hole (Figure 25), rather than walking up and down placing each shovelful into a random position that it did not initially come out of (Figure 26). Clearly for simple distributions of mass, like our potato trench, determining the optimal transportation plan is a trivial exercise, but for more complex distributions, this may not be as readily apparent. In simple regimes we can recognise a good solution, but for distributions of greater complexity, we must use a numeric quantifier to judge whether one method of filling in the trench is truly superior to another.

More effort must be exerted for each step that is taken, so it is clear that transport plans which force us to take more steps are inferior. This total work done may be used to judge whether we have found the best plan, as the optimal solution verifiably takes the smallest number of steps possible. This optimal value of work done may be interpreted as a measure of distance between two distributions, with smaller values of work done to transform one from another suggesting the two distributions possess a greater level of similarity.

For the 1-dimensional potato trench example, there exists a very fast method of determining the minimal quantity of total work done³⁰. If we know that we are forced into walking back and forth in a straight line, then an optimal solution may be obtained by keeping track of the demand/surplus of earth that is encountered as we make a single pass from one end of the path to the other. This process is best demonstrated with some examples, shown in Figures 27-29.

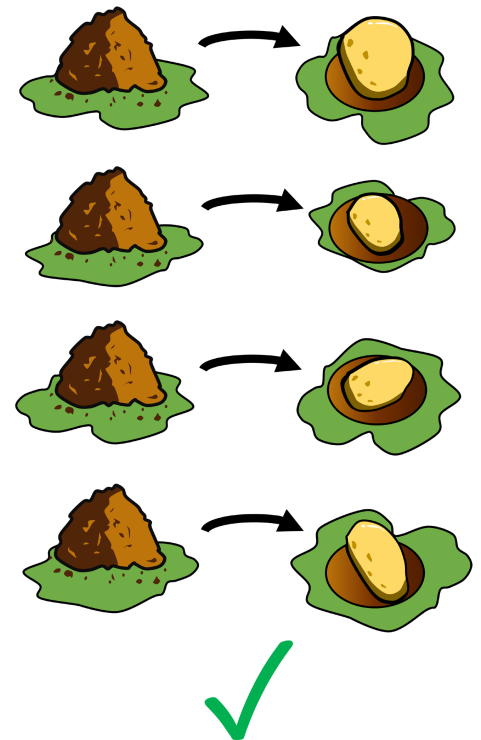


Figure 25: A simple example of the most efficient filling strategy for a dug out trench of potatoes.

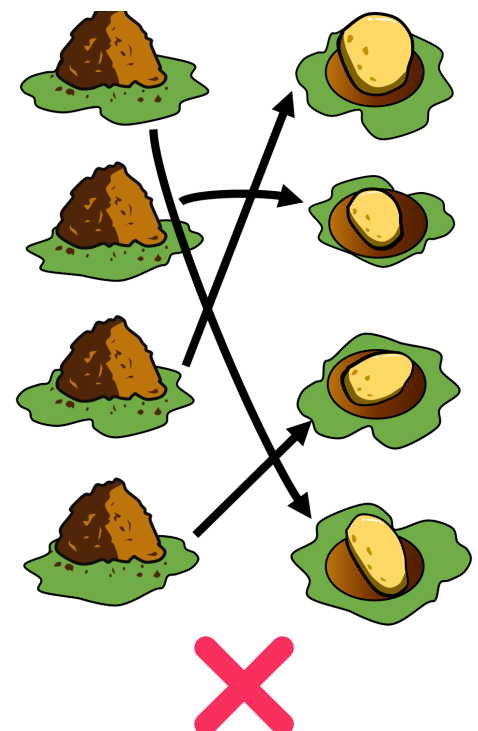
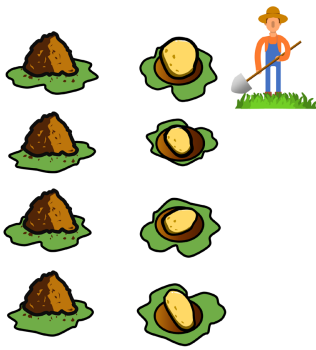
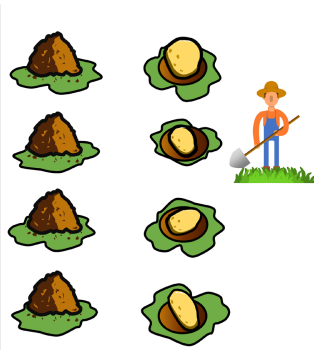


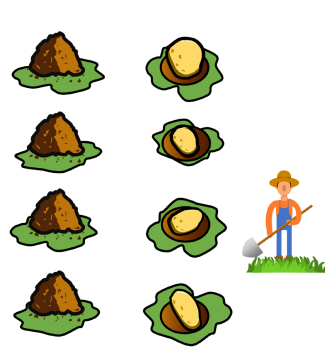
Figure 26: An inefficient method of transporting earth to each hole in the potato trench.



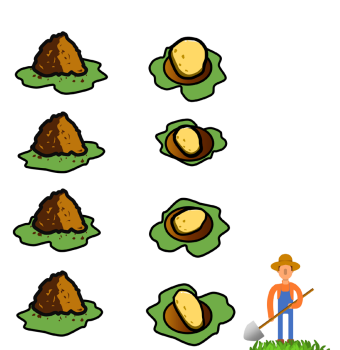
There is one hole which may be filled with the earth at this position so I will not need to increase the total work expended.



There is one hole which may be filled with the earth at this position so I will not need to increase the total work expended.



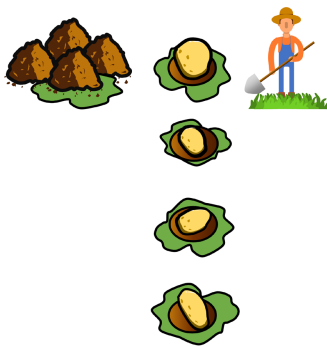
There is one hole which may be filled with the earth at this position so I will not need to increase the total work expended.



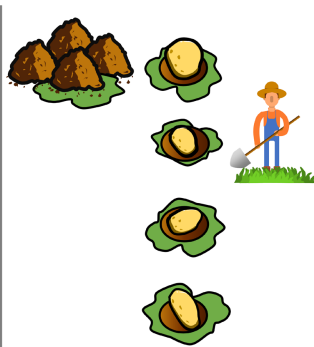
There is one hole which may be filled with the earth at this position so I will not need to increase the total work expended.

The total work required to move earth will thus be $0 + 0 + 0 + 0 = 0$

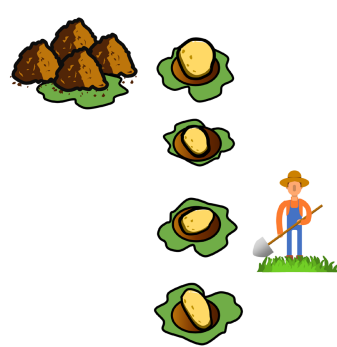
Figure 27: The calculation of EMD when two one-dimensional distributions are identical to one another.



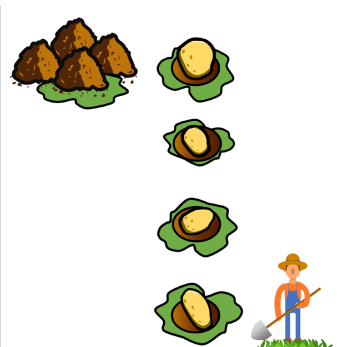
There is one hole which may be filled with the earth at this position so I will not need to increase the total work expended. I will remember there is an excess of three earth here.



There is one hole but no earth, I know there is surplus earth 1 meter away so I could fill this hole for a cost of 1.



There is one hole but no earth, I know there is surplus earth 2 meters away so I could fill this hole for a cost of 2.



There is one hole but no earth, I know there is surplus earth 3 meters away so I could fill this hole with a cost of 3.

The total cost will thus be $0 + 1 + 2 + 3 = 6$

Figure 28: The calculation of EMD if the first distribution of earth is in one pile.

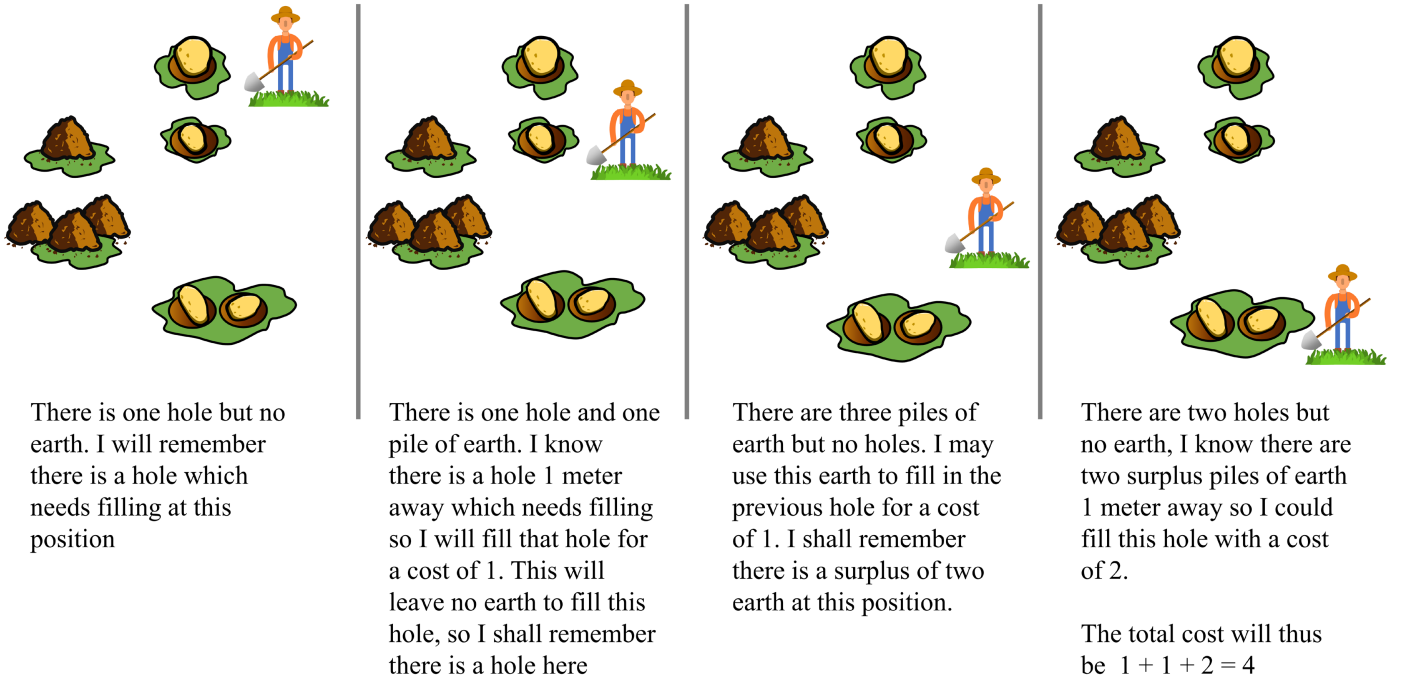


Figure 29: The calculation of EMD between two dissimilar distributions.

In each of these examples we know how many steps will need to be taken carrying earth after walking along the trench a single time. This process is very fast, and is applicable to any unbroken 1-dimensional distribution. For two monotonically spaced (where the n fixed positions mass can occupy have been spaced apart regularly) distributions X and Y this may be carried forward in four simple mathematical operations (Figure 30-31) by the equation:

$$EMD(X, Y) = \sum_{i=1}^n |CumulativeSum(X - Y)|$$

$X - Y$, we assume that the second distribution is initially positive, and thus we subtract this to signify that these are the holes that must be filled in at each position.

$CumulativeSum(X - Y)$, we take the cumulative sum by adding each successive vector element together as we sweep from left to right, keeping track of the cumulative total. We know work must

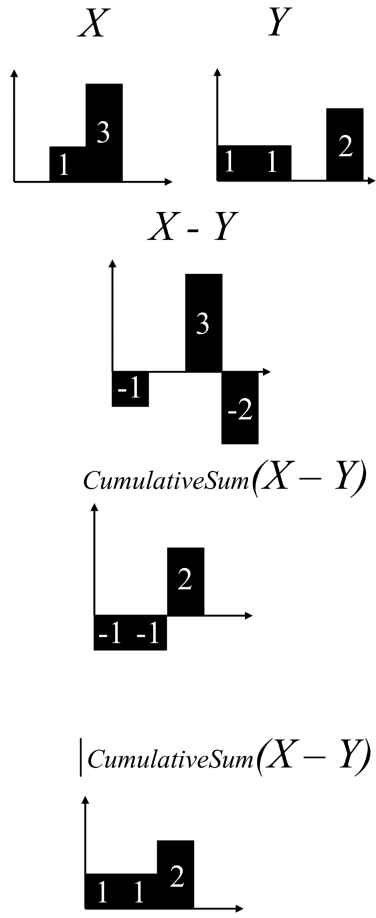


Figure 30: Each of the successive steps that are taken to transform two linear distributions when computing EMD.

$$\sum_{i=1}^n |CumulativeSum(X - Y)|$$

$$= 1 + 1 + 2 + 0$$

$$= 4$$

Figure 31: Summing each of the values in the computed vector gives a final value of EMD

be done at each step as we walk from one end of the path to the other, which is tracked by this operation.

$|CumulativeSum(X - Y)|$, the absolute values are taken as we know that walking back on ourselves to fill in a hole takes as much work as walking forward to fill in a hole.

$\sum_{i=1}^n |CumulativeSum(X - Y)|$, we take the summation across all of these values to give the total work that must be done at minimum to fill in the second distribution.

Unfortunately, there are many classes of transportation problems which do not fit the constraints that this method requires. The restriction to distributions which fall on unbroken monotonic 1-dimensional distributions means that if we were to dig holes across a 2-dimensional field depositing the earth in mounds at random (Figure 32), we would not be able to deduce the minimal cost of filling our holes back in via this method. Further, this algorithm does not return a transportation plan, which would be a useful piece of information for the person filling in the holes.

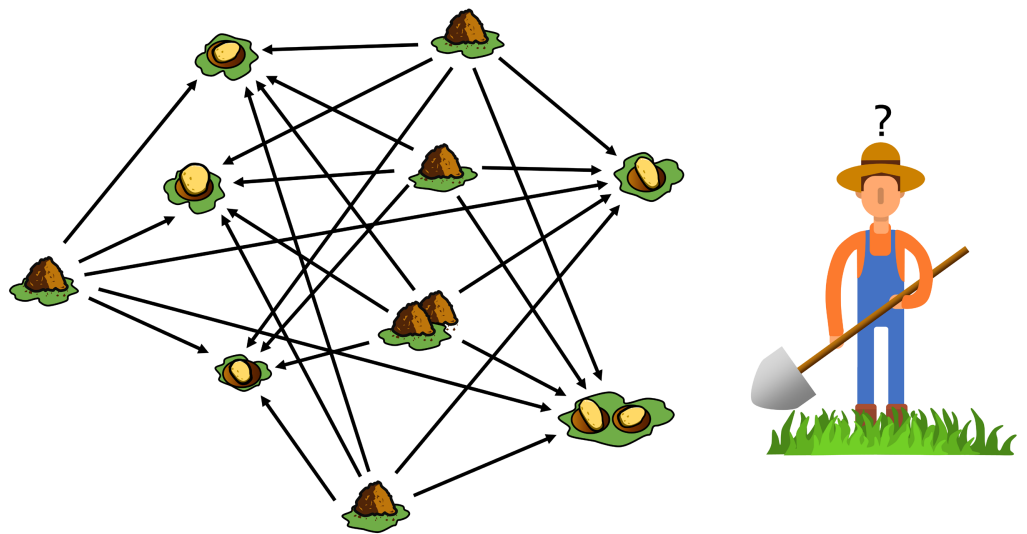


Figure 32: A 2-dimensional network of earth and holes with no clear method of choosing which hole to assign earth to.

Lots of algorithms for computing the optimal transportation plan for more complex spaces exist³¹. A network, or graph, may be formed by forming edges between each of the piles of earth and the holes (supply and demand nodes of a bipartite graph in network flow literature³²), ascribing the metric cost to transport earth between each node to each edge. These edge networks may be represented by square distance matrices, which represent the unit cost of transporting mass between each of the nodes. The mass of each node will be stored in one of two vectors, one for supply nodes and one for demand nodes. In recent years the Sinkhorn algorithm³³ has gained popularity in a range of applications³⁴ for the speed³⁵ at which it arrives at solutions, but in our work and others³⁶, we have found that this is outperformed by the network simplex algorithm³⁷. This is presumably due to the relatively small network sizes used in our work, and as such this is the approach taken in this thesis.

The simplex algorithm was developed roughly a decade before the network simplex and is one of the most robust and well-studied technique in optimisation. An in depth overview of the simplex algorithm is beyond the scope of this chapter, but suffice to say this provided strong motivation for the development of the network simplex algorithm. The simplex algorithm allows us to find the optimal values for a set of variables given a set of linear constraints and an objective function which must be maximised or minimised. By definition this focuses on spaces which can be constrained by linear equations.

One of the core theorems initially presented alongside the simplex method, was that any solvable linear optimisation problem must have a basic feasible solution. Basic in this instance means that all variables which can possibly be set to 0 are set to 0. All solvable problems will have at least one solution which is optimal and has at most as many non-zero variables as the number of constraints given by the initial problem. A linear problem with ten independent variables, and two constraints is guaranteed to have an

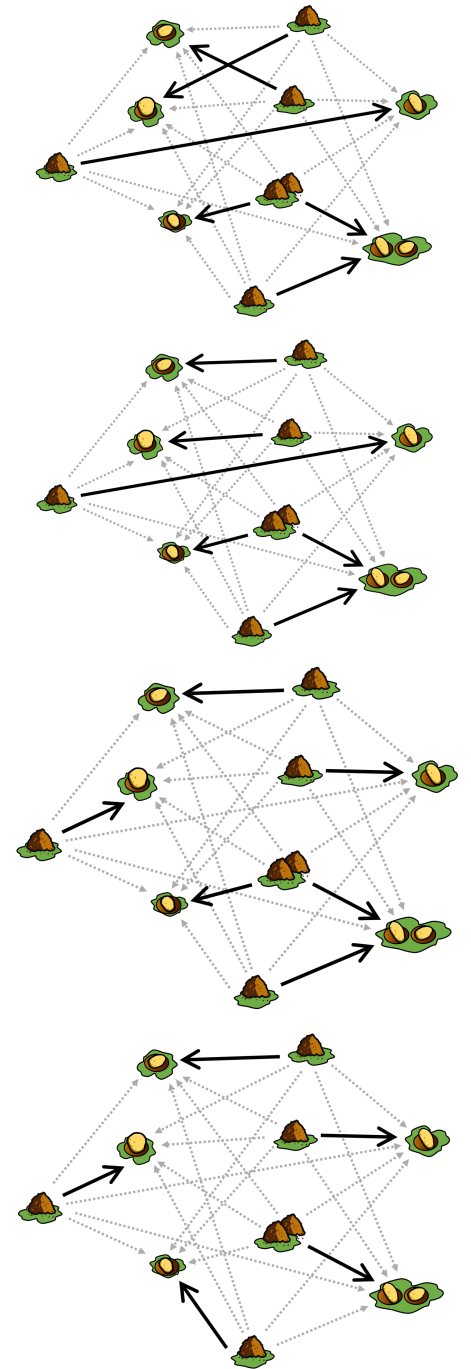
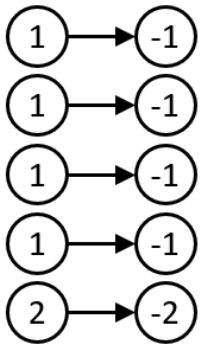
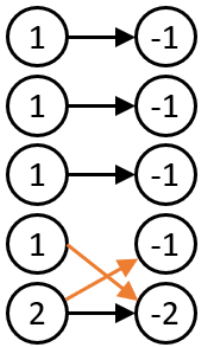
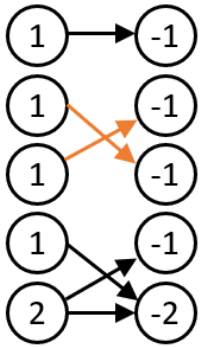
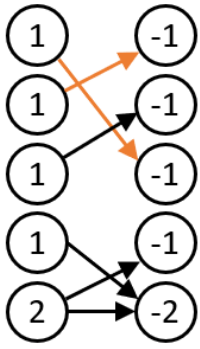


Figure 33: A sub-optimal feasible solution to fill holes with piles of earth (top) is successively improved by swapping edges so that the problem remains solved and the overall cost is reduced at each stage. Eventually, an optimal solution is found (bottom) where no further changes can reduce the total cost.



optimal solution where at least eight of the variables are equal to 0. The algorithm thus focuses on pivoting variables in and out of a smaller two variable basic solution until we have arrived at a solution which is both verifiably optimal and feasible.

For network (*i.e.* graph) problems, whilst the simplex algorithm can be applied to find an optimal matching, this is known to be a slow approach. The large number of linear constraints which must be used to represent the problem means that we must track a lot of variables at each step of the process. To alleviate this, the network simplex algorithm was proposed in 1976³⁸, with a provably efficient method given by Orlin in 1997³⁹, which remains one of the fastest methods in practice. The simplex method is adapted for networks, through the observation that the basic feasible solution in linear programming has a network analogue, the basic feasible spanning tree solution.

This states that there must exist an optimal solution to the network problem which contains the n nodes of the network, and at most $n-1$ edges connecting these nodes together. Therefore we may keep track of $n-1$ variables at each step of the process, setting all other edge assignments to be equal to 0. An initial solution to the problem is generated by choosing $n-1$ edges at random or through an initialisation scheme. This edge assignment may not be a feasible solution to the problem, *i.e.* some holes may be unfilled, or mass assigned to too many places. Some edges may appear in the solution with zero mass assigned to them to ensure a spanning tree solution. We choose edges to pivot in and out of the spanning tree until we have reached a point where the solution is both feasible, and swapping out two edges will not lead to a smaller total cost to the problem. At this point we have arrived at an optimal spanning tree solution, which is provably minimal. Astute readers may have noticed that this is a disconnected graph, and thus not a tree. This nomenclature stems from network flow problems which assume that all mass is sourced from a single node, however this is still a valid transport plan. This plan may be returned to the user, and the

Figure 34: The same sequence of steps taken in Figure 33, but demonstrated on a bipartite graph. Masses are overlaid on each node with positive values indicating a source, and negative values a sink, with the swapping edges at each stage given in orange.

total work done calculated by multiplying the total mass transported along each edge by the cost of each edge.

Given the ubiquity of this problem across so many fields, it has been discovered and rediscovered multiple times, giving it a rich and fragmented history. The earliest documented approach to solving this problem was formulated by Gaspard Monge in 1781⁴¹ to find the best routes sending construction and military supplies across Napoleonic France. As a guarded national secret this work was not publicised during Monge's lifetime, and the technique was not developed much further. Over a century later the Russian mathematician Leonid Kantorovich saw how the optimal transport algorithm presented by Monge would be improved by the latest advances in linear programming, and in 1942 went on to develop a much faster approach⁴². Due to its potential for exploiting economic markets, this method was deemed antithetical to soviet doctrine and Kantorovich was dissuaded from discussing his work by the local authorities, with this same approach independently found by Georg Dantzig⁴³.

A Russian statistical physicist, Leonid Wasserstein, went on to use the technique to describe Markov processes of large systems of automata. The director of his institute, Roland Dobrushin, being impressed with this idea, promoted his colleagues work as the Wasserstein metric, which remains a commonly used term in statistics and applied mathematics to this day⁴⁴. At the turn of the century, researchers at MIT applied this method to judge the similarity between images by looking at the optimal transport plan between binned histogram representations⁴⁵. Rather than referring to this as the complete weighted bipartite minimal cost matching problem, how this class of problems may be called within computer science, they renamed the technique the Earth Movers Distance (EMD).

In the past decade, an approximation method for the optimal transport plan called the Sinkhorn algorithm has been formulated

³⁵, which allows the transport plan to be computed in a constant number of iterations within a provably tight margin of error, and has been referred to as the Sinkhorn distance. Cédric Villani won his Fields medal applying optimal transport measures to gas equations, and has opined that if the technique is not referred to as optimal transport, then it should be named the Monge-Kantorovich distance, after those who originally formulated the solution ⁴⁴. In this work, we have chosen to continue referring to these optimal transport measures as the earth movers distance due to the immediacy of understanding this description conveys to a wide audience.

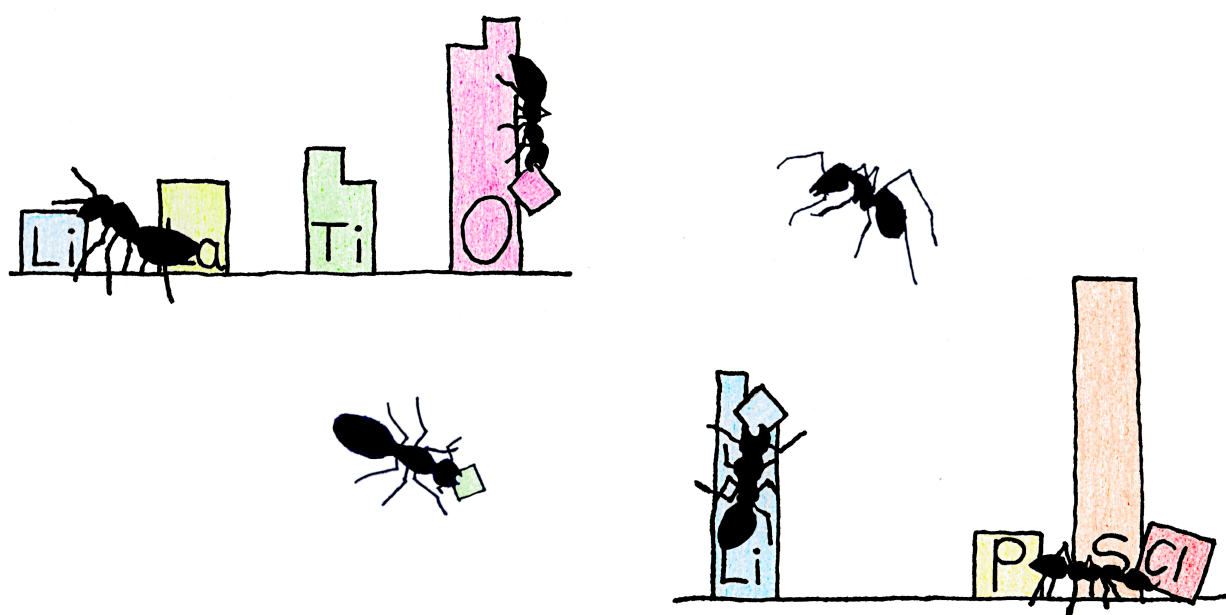
References

1. Chen, A., Zhang, X. & Zhou, Z. Machine learning: Accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).
2. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Acc. Mater. Res.* **3**, 597–607 (2022).
3. Vela, S., Laplaza, R., Cho, Y. & Corminboeuf, C. cell2mol: encoding chemistry to interpret crystallographic data. *Npj Comput. Mater.* **8**, 1–8 (2022).
4. Bostrum, Nick. *Superintelligence: Paths, Dangers, Strategies*. (Oxford University Press, 2014).
5. Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *Npj Comput. Mater.* **8**, 1–26 (2022).
6. Sendek, A. D. *et al.* Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
7. Crabtree, G. The coming electric vehicle transformation. *Science* **366**, 422–424 (2019).
8. Keck, F., Lenzen, M., Vassallo, A. & Li, M. The impact of battery energy storage for renewable energy power grids in Australia. *Energy* **173**, 647–657 (2019).
9. Wang, Y. *et al.* Lithium and lithium ion batteries for applications in microelectronic devices: A review. *J. Power Sources* **286**, 330–345 (2015).
10. Wu, W. *et al.* A critical review of battery thermal performance and liquid based battery thermal management. *Energy Convers. Manag.* **182**, 262–281 (2019).
11. Wang, Y. *et al.* Design principles for solid-state lithium superionic conductors. *Nat. Mater.* **14**, 1026–1031 (2015).

12. Tikekar, M. D., Choudhury, S., Tu, Z. & Archer, L. A. Design principles for electrolytes and interfaces for stable lithium-metal batteries. *Nat. Energy* **1**, 1–7 (2016).
13. Hyooma, H. & Hayashi, K. Crystal structures of $\text{La}_3\text{Li}_5\text{M}_2\text{O}_{12}$ ($\text{M}=\text{Nb}, \text{Ta}$). *Mater. Res. Bull.* **23**, 1399–1407 (1988).
14. Thangadurai, V., Kaack, H. & Weppner, W. J. F. Novel Fast Lithium Ion Conduction in Garnet-Type $\text{Li}_5\text{La}_3\text{M}_2\text{O}_{12}$ ($\text{M}=\text{Nb}, \text{Ta}$). *J. Am. Ceram. Soc.* **86**, 437–440 (2003).
15. Buschmann, H. *et al.* Structure and dynamics of the fast lithium ion conductor ‘ $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$ ’. *Phys. Chem. Chem. Phys. PCCP* **13**, 19378–19392 (2011).
16. Kobayashi, S. *et al.* Lithium Lanthanum Titanate Single Crystals: Dependence of Lithium-Ion Conductivity on Crystal Domain Orientation. *Nano Lett.* **22**, 5516–5522 (2022).
17. Hu, Z. *et al.* Enhanced Li ion conductivity in Ge-doped $\text{Li}_{0.33}\text{La}_{0.56}\text{TiO}_3$ perovskite solid electrolytes for all-solid-state Li-ion batteries. *New J. Chem.* **42**, 9074–9079 (2018).
18. Bruce, P. G. & West, A. R. The A-C Conductivity of Polycrystalline LISICON, $\text{Li}_2 + 2x \text{Zn}_{1-x} \text{GeO}_4$, and a Model for Intergranular Constriction Resistances. *J. Electrochem. Soc.* **130**, 662 (1983).
19. Hu, Y.-W., Raistrick, I. D. & Huggins, R. A. Ionic Conductivity of Lithium Orthosilicate—Lithium Phosphate Solid Solutions. *J. Electrochem. Soc.* **124**, 1240 (1977).
20. Bates, J. B. *et al.* Fabrication and characterization of amorphous lithium electrolyte thin films and rechargeable thin-film batteries. *J. Power Sources* **43**, 103–110 (1993).
21. Goodenough, J. B., Hong, H. Y.-P. & Kafalas, J. A. Fast Na^+ -ion transport in skeleton structures. *Mater. Res. Bull.* **11**, 203–220 (1976).
22. Aono, H., Sugimoto, E., Sadaoka, Y., Imanaka, N. & Adachi, G. Ionic Conductivity of Solid Electrolytes Based on Lithium Titanium Phosphate. *J. Electrochem. Soc.* **137**, 1023 (1990).
23. Xu, X., Wen, Z., Wu, X., Yang, X. & Gu, Z. Lithium Ion-Conducting Glass–Ceramics of $\text{Li}_{1.5}\text{Al}_{0.5}\text{Ge}_{1.5}(\text{PO}_4)_3-x\text{Li}_2\text{O}$ ($x=0.0-0.20$) with Good Electrical and Electrochemical Properties. *J. Am. Ceram. Soc.* **90**, 2802–2806 (2007).
24. Kamaya, N. *et al.* A lithium superionic conductor. *Nat. Mater.* **10**, 682–686 (2011).
25. Deiseroth, H.-J. *et al.* $\text{Li}_6\text{PS}_5\text{X}$: A Class of Crystalline Li-Rich Solids With an Unusually High Li^+ Mobility. *Angew. Chem. Int. Ed.* **47**, 755–758 (2008).
26. Snir, A. *et al.* The Origin of Cultivation and Proto-Weeds, Long Before Neolithic Farming. *PLoS ONE* **10**, e0131422 (2015).
27. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).

28. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
29. Widdowson, D. & Kurlin, V. Pointwise distance distributions of periodic point sets. Preprint at <https://doi.org/10.48550/arXiv.2108.04798> (2022).
30. Kolouri, S., Pope, P. E., Martin, C. E. & Rohde, G. K. Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model. Preprint at <http://arxiv.org/abs/1804.01947> (2018).
31. Flamary, R. *et al.* POT: Python Optimal Transport. 8.
32. Ahuja, Ravindra K., Thomas L. Magnanti, & James B. Orlin. Network flows. (1988)
33. Sinkhorn, R. A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices. *Ann. Math. Stat.* **35**, 876–879 (1964).
34. Wong, E., Schmidt, F. & Kolter, Z. Wasserstein Adversarial Examples via Projected Sinkhorn Iterations. in *Proceedings of the 36th International Conference on Machine Learning* 6808–6817 (PMLR, 2019).
35. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. in *Advances in Neural Information Processing Systems* vol. 26 (Curran Associates, Inc., 2013).
36. Dong, Y., Gao, Y., Peng, R., Razenshteyn, I. & Sawlani, S. A Study of Performance of Optimal Transport. Preprint at <http://arxiv.org/abs/2005.01182> (2020).
37. Ahuja, R., Magnanti, T. & Orlin, J. *Network Flows: Theory, Algorithms, and Applications*. (Pearson, 1993).
38. Cunningham, W. H. A network simplex method. *Math. Program. Ser. B* **11**, 105–116 (1976).
39. Orlin, J. B. A polynomial time primal network simplex algorithm for minimum cost flows. *Math. Program.* **78**, 109–129 (1997).
40. Carrière, M., Cuturi, M. & Oudot, S. Sliced Wasserstein kernel for persistence diagrams. in *Proceedings of the 34th International Conference on Machine Learning*. **70**, 664–673 (2017).
41. Monge, G. Mémoire sur la théorie des déblais et des remblais *Histoire de l’Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*. 666-704 (1781).
42. Kantorovich, L. On the translocation of masses. *C.R. (Doklady) Acad. Sci. URSS (N.S.)*. **37**, 199–201. (1942).
43. Fulkerson, D. R., Ray, D. & Dantzig G. B. Computation of Maximal Flows in Networks (1955).
44. Villani, C. Optimal Transport: Old and New, Berlin: Springer (2009).
45. Rubner, Y., Tomasi, C. & Guibas, J. G. The Earth Movers Distance as a Metric for Image Retrieval. *Int J Comput Vis.* **40**, 999-121 (2000).

The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions



This work was accepted for publication in *Chemistry of Materials*, 24, 32, 2020. I was the primary author, and performed each of the calculations including all development of the Element Movers Distance (ElMD) code base. This work was co-authored and supervised by Matthew S. Dyer, Michael W. Gaultois, Vitaliy A. Kurlin, and Matthew J. Rosseinsky.

Reprinted with permission from *Chem. Mater.* 2020, 32, 24, 10610-10620. Copyright 2020 American Chemical Society.

Introduction

Even before Aristotle, philosophers sought to explain the properties of materials through their elemental composition. As an experimental chemist, the first step in any investigation is choosing

what elements in what ratio to put into the sample, and the composition is arguably the most important independent variable under control. In many functional materials, where disorder is important to the functional properties (such as electronic or ionic conductivity), the elemental composition is an invariant property that is well described. This is because the nominal composition that is put into a synthetic process is generally well-defined, and also because there are extensive characterization methods to experimentally determine the elemental composition.

Although the underlying theory has evolved considerably since antiquity, the elemental composition of a material continues to be a prime director of materials properties, and we know now the chemical composition largely dictates the nature of the chemical bonding, which has a strong influence on the crystal structure and physical properties. Similar compositions lead to similar properties, and when estimating material properties, it is important to consider the closest known composition to the one being considered. These similarities can be defined quantitatively in a distance function which returns a real valued number, such that identical objects have a distance of 0, and less similar objects return a larger value. We would expect that small changes in chemical makeup would lead to correspondingly small variations in chemical property, and that chemically dissimilar compounds may behave entirely differently.

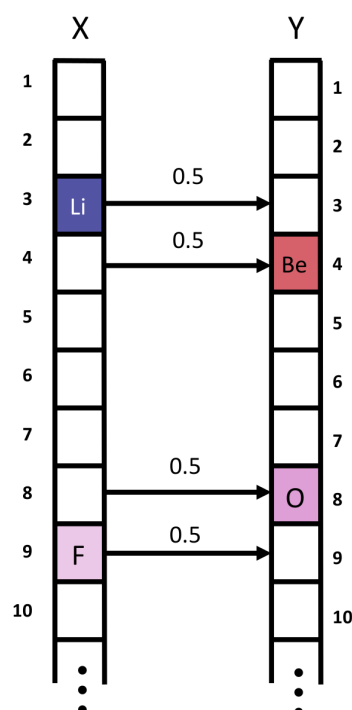
The chemist develops such understanding naturally through their exploration of the sciences. Whilst each practitioner may have a personal tolerance for what they believe to be “chemically similar”, two compositions which differ only by a minor dopant or by the substitution of a similar element have inarguable similarity. This relationship may be immediately clear to the chemist, but in practice it is difficult to capture these small physical changes numerically. In this chapter we present a novel technique for calculating the distance between two compositions, which captures nuanced variations in stoichiometry for both similar and dissimilar

compounds using our chemically inspired implementation of the Earth Movers Distance, the Element Movers Distance (EIMD).

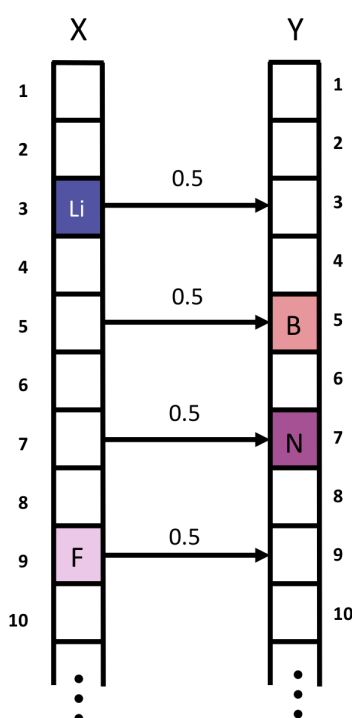
By correlating relationships between the chemical composition of materials and their observed behaviour, we may detect underlying statistical relationships between these. This can be used in an automated process to usefully inform the chemistry, whether it be by providing the relationships with other clusters of similar materials, or even by estimating the properties. This has been exploited implicitly by modern machine learning (ML) methods, which have been applied to capitalize on the strong determination of properties by composition; there are many reports of regression models to estimate material performance of inorganic solids from composition alone¹⁻³.

For these models to be successful we require two things: a large collection of data, and a method of differentiating these such that we may uncover the subtle relationships which govern a material's properties. Having a metric to quantify these relationships allows us to take our bearings and construct maps of chemical space to enable clear exploration, providing an awareness of compositional relationships between materials. When predicting the properties of a new composition, we must form an understanding of its relation to other reported compounds, and the distinguishing quality of the similarity metric chosen can vastly affect performance. The choice of metric should therefore possess enough fidelity to give an accurate representation of chemical relationships between entries in a database of compositions, and align with human understanding.

Though widely used as a metric, the compositional Euclidean distance (CED) can perform poorly at the task of distinguishing compounds. A common method of encoding a composition is to store the relative ratio of each element in the compound to its associated index in a vector of length 103, for each of the naturally stable elements. Taking x_i to be the fraction of the i -th element in a



LiF \rightarrow BeO = 1



LiF \rightarrow BN = 1

compound X, we take the CED to a second such vector, Y, via the standard formula:

$$\sqrt{\sum_i (x_i - y_i)^2}$$

Due to the sparsity of these vectors, the CED overly simplifies and exaggerates physical differences. As an example by taking the atomic number of the 103 stable elements as our index, the compositional vectors of LiF and BeO would be $[0_0 \dots 0.5_3 \dots 0.5_9 \dots 0_{103}]$ and $[0_0 \dots 0.5_4 \dots 0.5_8 \dots 0_{103}]$ respectively. Taking the non-zero elements with indices 3, 4, 8, and 9, the CED between these vectors would thus be:

$$\sqrt{(0.5 - 0)^2 + (0 - 0.5)^2 + (0 - 0.5)^2 + (0.5 - 0)^2} = \sqrt{1}$$

A third binary composition, BN, which with a compositional vector of $[0_0 \dots 0.5_5 \dots 0.5_7 \dots 0_{103}]$ is arguably less chemically similar to LiF than BeO, would also have a CED of 1 to both of these compounds, as demonstrated in Figure 1. A CED of 1 would be calculated between any two binary compositions which do not have a common element. This discrete nature of the CED does not provide an accurate distinction between compounds which may be entirely different chemically, and while this may capture local trends in a chemical dataset, will lose global information. We can improve on this shortcoming by incorporating a measure of elemental similarity which may be applied to a compositional vector directly.

The Earth Mover's Distance (EMD) is a metric which is well-constructed to pair elements between compositions, and from this judge their similarity, which has had successful applications in multiple fields⁴⁻⁶. The EMD may analogously be thought of as the minimal amount of work to move piles of earth to fill pits of equal overall volume but different shapes, a long studied transportation problem⁷ with fast algorithmic implementations⁸⁻¹¹ that are discussed in chapter 1. This consistently returns a unitless quantity of work which may be interpreted as a measure of distance.

Figure 1: The CED is demonstrated between compositional vectors of LiF, BeO, and BN by taking the absolute difference of atomic sites, with the atomic number as index.

We could assign the atomic number as the vector index for each element, then take the difference between indices as a measure of elemental similarity, but this approach loses the natural clustering of chemical properties afforded by the periodic table. An ideal elemental indexing would perfectly capture the chemical trends observed in nature, but ordering the elements in such a manner is problematic. As well as the unclear resolution of how to handle the f-block elements, chemical trends moving down the periodic table tend to be the direct opposite of those moving across. This leads to some elements having greater substitutional feasibility to their diagonal neighbour than their immediate neighbour, making a simple placement of these difficult.

To solve this problem, Pettifor proposed a method of labelling the elemental scale in his seminal paper of 1984¹², drawn from extensive domain knowledge. These numeric labels may form the basis of a coordinate system allowing us to associate patterns in geometric and physiochemical properties, with extensions to this idea continuing to guide practitioners^{13,14}. This concept of labelling was further developed by analysing the probability that an element can be substituted for another given the same structural framework, on 20,500 compounds of the ICSD by Glawe et al.¹⁵. This probability matrix can be reordered to maximize the likelihood that local neighbourhoods will contain elements with greater feasibility of stable substitutions, thus possessing inherent chemical similarities¹⁶. We take the associated indices of this final ordering to give each element its modified Pettifor number, as demonstrated in Figure 2.

In this chapter we define a composition vector by taking the ratio of each element in a compound assigned to the index of its respective modified Pettifor number. By assuming the sample of the set of feasibly stable compounds (although we know this is not strictly the case¹⁷), we can see that these indices capture the truly physical similarities between elements from statistical analysis, which has been implemented in the Element Movers Distance

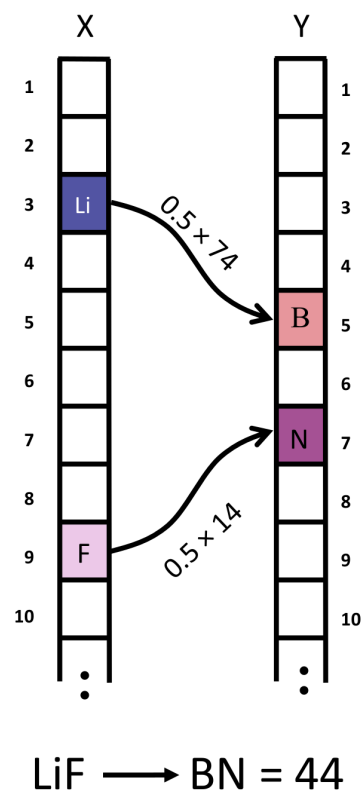
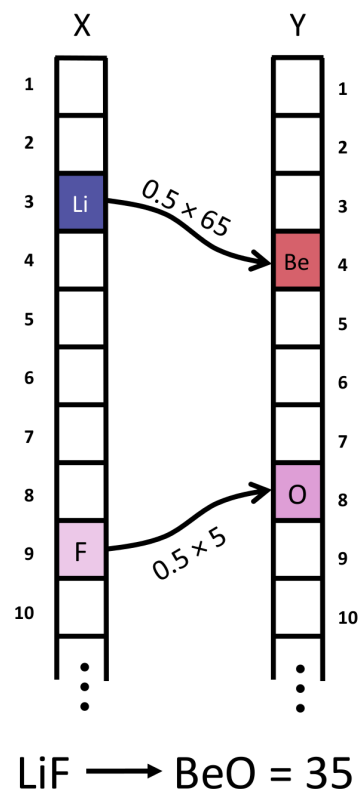


Figure 2: The equivalent EIMD solution to Figure 1, where instead elements are matched to one another by similarity. By calculating the cost to transport each atom along the modified Pettifor scale we have arrived at a distance which is reflective of the chemical dissimilarity.

(EIMD) library. Using the modified Pettifor scale gives resultant similarities between compounds which align with human judgement, but may be substituted with any continuous elemental scale including unequally spaced distributions such as Pauling electronegativity or high dimensional machine learnt elemental vector embeddings.

This gives us the ability to place new compositions within the context of previously reported compounds allowing us to attribute properties to these before the lengthy process of synthesis. We can do this automatically with ML techniques, where the EMD forms part of the workflow to predict properties quantitatively. We may additionally assign properties to compositions qualitatively, by simply searching through multiple databases to find the most similar existing entries. This second approach requires the practitioner’s judgement on whether to take the property of the closest match, an average of many similar compounds, or to conclude that the reported landscape is not sufficiently complete to make an accurate judgement.

The Earth Mover’s Distance

We take an initial matching by pairing each of the m elements in a vector, X , to its most similar unmatched partner in the n elements of a second vector Y , until all have been paired. The parameters we alter is the transportation plan which is the quantity matched, q , from the i -th element of X , to the j -th element of Y , given by q_{ij} . A cost is calculated (Eq. 1a) by summing all quantity, q , paired through each matching multiplied by the difference in indices on the modified Pettifor scale, p , between the elements matched. When all elements have been paired, this is a feasible solution to the problem, however it may not be an optimally minimal solution. Given two vectors we take a feasible matching, and successively improve this via the network simplex algorithm until the total summed cost is verified optimal. For a compositional vector X , $\sum_{i=1}^m x_i = 1$ and therefore the total quantity matched with any other vector will therefore also equal 1. As we are describing a

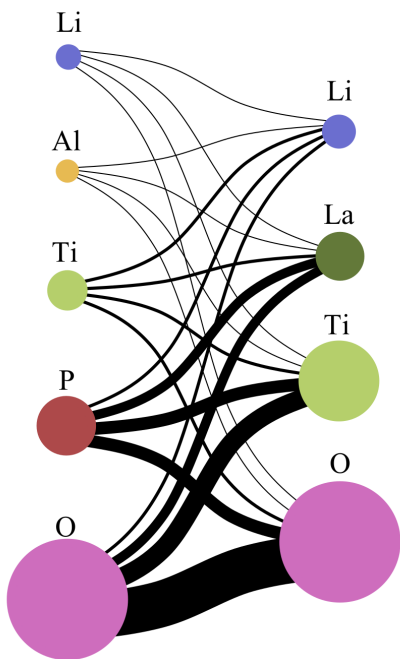


Figure 3 All feasible solutions to compute the EMD between $Li_{1.3}Al_{0.3}Ti_{1.7}(PO_4)_3$ and $La_{0.5}Li_{0.35}TiO_3$. The area of each disk represents the fraction of the element in the compound, and the width of each arc shows the maximal quantity which could be theoretically matched within the constraints of the problem. Figure not to scale.

transformation from two distributions with the same integral, this satisfies the triangle inequality which is proven in Appendix A of reference 4. The formal definition of the EMD between two compositional vectors $X = (x_1 \dots x_m)$ and $Y = (y_1 \dots y_n)$, as given in reference 18, Eq 1a, is defined by the optimisation problem:

$$EMD(X, Y) = \min_{q_{ij}} \sum_{i=1}^m \sum_{j=1}^n q_{ij} |p_i - p_j| \quad (1a)$$

$$\text{subject to } q_{ij} \geq 0 \text{ for any } i, j \quad (1b)$$

$$\sum_{j=1}^n q_{ij} \leq x_i \text{ for any } 1 \leq i \leq m \quad (1c)$$

$$\sum_{i=1}^m q_{ij} \leq y_j \text{ for any } 1 \leq j \leq n \quad (1d)$$

$$\sum_{i=1}^m \sum_{j=1}^n q_{ij} = 1 \quad (1e)$$

Constraint (1b) defines that we may only match a positive quantity from X to Y , (1c) and (1d) state that each element will only pair to another up to its ratio. The final constraint ensures that all of the elements in X are matched to an element in Y such that a feasible solution has been achieved. Note that whilst the total EMD cost is an optimally minimal value, a problem can have multiple transportation plans that lead to the same optimal solution.

Taking three candidate solid state electrolytes with known dissimilarity in composition and structure to exemplify this as in Figure 4, we can see how the EIMD allows greater depth of analysis when defining chemical similarity compared to the CED. From Figure 4 we see how the solution not only gives us the

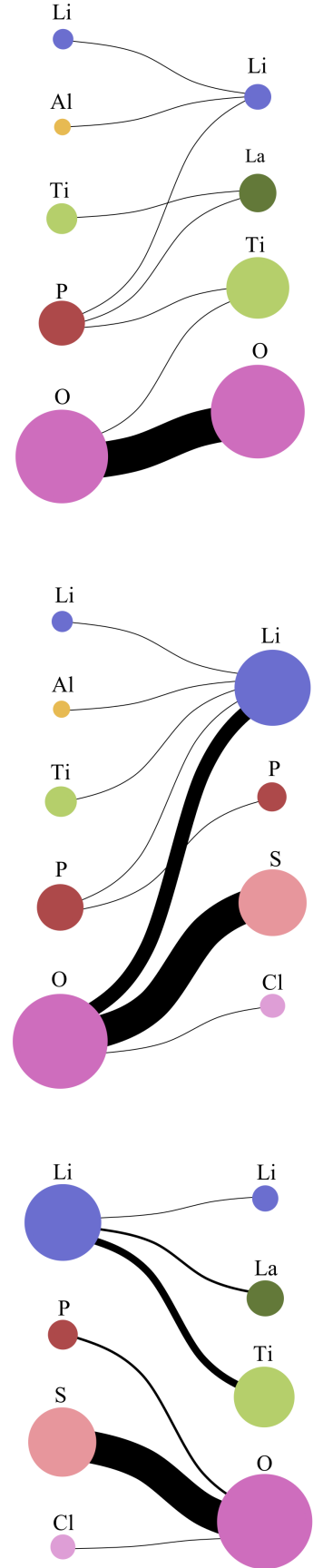


Figure 4: The optimal matchings between $Li_{1.3}Al_{0.3}Ti_{1.7}(PO_4)_3$ and $La_{0.5}Li_{0.35}TiO_3$ (EIMD=12.70), $Li_{1.3}Al_{0.3}Ti_{1.7}(PO_4)_3$ and Li_6PS_5Cl (EIMD=31.08), and $La_{0.5}Li_{0.35}TiO_3$ (EIMD=18.38)

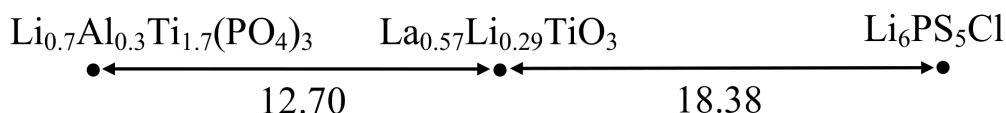


Figure 5: A simple embedding of $\text{Li}_{1.3}\text{Al}_{0.3}\text{Ti}_{1.7}(\text{PO}_4)_3$, $\text{La}_{0.5}\text{Li}_{0.35}\text{TiO}_3$, and $\text{Li}_6\text{PS}_5\text{Cl}$ using calculated EIMD distances

measure of distance, but also the quantity of elements that are paired to one another. These resultant distances may be used to define a geometry on the compositional space, as in Figure 5. In this instance the three compositions fall on a geodesic, or straight line. We may apply this metric to any two chemical formulae, enabling us to highlight chemically similar substitutions, and thus familial relation, which may not have been immediately obvious from the compound formula. The reference implementation of the Element Movers Distance may be found at:

<https://github.com/lrcfmd/EIMD/>.

Pairing Structures to Compositions

It is widely recognized that composition is not the sole determinant of physical performance, and crystal structure plays a fundamental role in property, which can be dependent on many different length scales. Codifying these structures such that we may compare them for similarity has known difficulties¹⁹ due to the periodicity of inorganic systems. For organic molecules there exist methods of formally encoding a structure derived from the strict lexicographic conventions of organic chemistry^{20,21}, and methods of encoding an inorganic crystals local environment and symmetry have been successfully implemented^{22–24}. Structural features are a known asset to ML models, and the addition of this information generally gives stronger predictive performance at screening and property prediction^{25–28}.

Unfortunately, structural information is often not reported in tandem with experimentally determined chemophysical properties, and many such properties are reported from solid solutions where a similar reported structure may not even exist. In many cases only the composition and the property under investigation will be

reported, leaving a fragmented data landscape with a barrier between databases. With the number of reported compounds untenable for any person to feasibly audit, we must bring this information together in an automated manner. The ElMD allows us to connect compounds to their closest determined structure in such a fashion, allowing databases with compositional information to be joined to databases of structural information.

Using the ElMD we may pair query formulae, including those which have never been synthesized, to their most similar compositions in one of the many chemical databases such as the Inorganic Crystal Structure Database (ICSD, 2017) ²⁹, consisting of 188,631 crystalline information files (cifs). A recent review of materials with reported ionic conductivity was undertaken with 842 compounds identified, which shall be discussed in greater detail in the next chapter. Each compound had a brute force comparative search applied to every ICSD entry, and the most similar pairings analysed by a team of 21 researchers at the Materials Innovation Factory, University of Liverpool, with the quality of these matchings assessed.

Of these compounds, 528 had a perfect match to a cif or an exact match under a minor change in stoichiometry. A further 254 compounds having a matching cif with a small number of elemental substitutions and similar in crystal structure. The remaining 60 formulae did not find a good match, mostly due to the materials being reported more recently than any database entries. In Table 1 we see some commonly cited compounds from this field and their closest matches in the ICSD. Clearly a distance of zero gives an exact match barring polymorphs, however there remains the more general case where an exact structure has not been reported. We can see that in each example a chemically similar compound has been returned, and we would expect the extracted structural features to have a high degree of correlation with the true structural information.

Table 1: The top three most similar results when querying some commonly cited solid state electrolytes against the ICSD (2017) with EIMD. Whilst there are only four queries with exact matches due to the recency in the electrolytes reporting, it may be seen that the remainder are chemically similar.

| Query | Three Closest Matches | EIMD |
|--|---|-------|
| $\text{Li}_{1.3}\text{Al}_{0.3}\text{Ti}_{1.7}(\text{PO}_4)_3$ | $\text{Na}_{1.261}\text{Al}_{0.302}\text{Ti}_{1.696}(\text{PO}_4)_3$ | 0.231 |
| | $\text{Li}_{1.2}\text{Al}_{0.2}\text{Ti}_{1.8}(\text{PO}_4)_3$ | 0.302 |
| | $\text{Li}_{1.4}\text{Al}_{0.4}\text{Ge}_{0.4}\text{Ti}_{1.2}(\text{PO}_4)_3$ | 0.368 |
| $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ | $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ | 0.000 |
| | $\text{Li}_{10}\text{SnP}_2\text{S}_{12}$ | 0.040 |
| | $\text{Li}_{9.81}\text{Sn}_{0.81}\text{P}_{2.19}\text{S}_{12}$ | 0.439 |
| $\text{Li}_6\text{PS}_5\text{Cl}$ | $\text{Li}_6\text{AsS}_5\text{I}$ | 0.231 |
| | $\text{Li}_6\text{PO}_5\text{Cl}$ | 0.385 |
| | $\text{Li}_6\text{PO}_5\text{Br}$ | 0.462 |
| $\text{Li}_5\text{La}_3\text{Nb}_2\text{O}_{12}$ | $\text{Li}_5\text{La}_3\text{Nb}_2\text{O}_{12}$ | 0.000 |
| | $\text{Li}_5\text{La}_3\text{Ta}_2\text{O}_{12}$ | 0.091 |
| | $\text{Li}_{5.08}\text{La}_3\text{Ta}_{1.51}\text{Zr}_{0.39}\text{O}_{12}$ | 0.322 |
| $\text{Li}_{1.5}\text{Al}_{0.5}\text{Ge}_{1.5}(\text{PO}_4)_3$ | $\text{Na}_{1.5}\text{Sn}_{1.5}\text{Sb}_{0.3}(\text{PO}_4)_3$ | 0.543 |
| | $\text{BaGa}(\text{PO}_4)_2$ | 0.712 |
| | $\text{BaSn}(\text{PO}_4)_2$ | 0.745 |
| $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$ | $\text{Li}_7\text{La}_3\text{Zr}_2\text{O}_{12}$ | 0.000 |
| | $\text{Li}_7\text{La}_3\text{Hf}_2\text{O}_{12}$ | 0.083 |
| | $\text{Li}_{7.1}\text{La}_3(\text{Zr}_{1.9}\text{Cr}_{0.1})\text{O}_{12}$ | 0.200 |
| $\text{Li}_{14}\text{Zn}(\text{GeO}_4)_4$ | $\text{Li}_{14}\text{Zn}(\text{GeO}_4)_4$ | 0.000 |
| | $\text{Li}_6\text{Ge}_2\text{O}_7$ | 0.410 |
| | $\text{Li}_6(\text{Si}_2\text{O}_7)$ | 0.543 |

Overall, 8.43% of the compounds were poorly matched as judged by a team of experimental chemists, although the number of false matches can be reduced by filtering the dataset with the application of a maximum threshold value. By removing all matches which have a distance greater than 1, we discard 12% of the entries, but improve the false positive rate to 5.7%. Although caution should be applied from the introduction of known errors, this provides a simple method for the automated creation of datasets on the scale required to uncover complex statistical relationships.

By assessing those matchings in Table 1 that are imperfect, we may see how the top ranked match remains structurally related,

with the remainder being simple dopings and substitutions. Of interest is $\text{Li}_{1.5}\text{Al}_{0.5}\text{Ge}_{1.5}(\text{PO}_4)_3$ in row 5, where whilst there is clear chemical relation between these three compounds, we see some dissimilarity in their structures, as $\text{Na}_{1.5}\text{Sn}_{1.5}\text{Sb}_{0.3}(\text{PO}_4)_3$, $\text{BaGa}(\text{PO}_4)$, and $\text{BaSn}(\text{PO}_4)$ represent the NASICON and pyrosilicate phases respectively.

It is likely, perhaps certain, that the chemist develops a deeper understanding of the relations between the compounds they study, their combinations, and their behaviours under environmental conditions than can be captured by a simple number. An engineered representation of compositions has however allowed us to express chemical relationships which have not previously been possible to express quantifiably. With a clearly defined metric of chemical similarity, we may use this as we would any other distance, with additional confidence that the underlying mechanics are mathematically aligned with chemical knowledge.

Mapping Compositional Space

The discovery of new materials has always been data driven, and mapping compositions to predict the existence of structures is a time honoured technique in crystallography^{30,31}. The visual medium provides a tangible clarity to the human reader, where abstract relations between compounds can be difficult to conceptualize through numerical analysis. The EIMD in conjunction with modern visualization techniques has clear application in this regard, giving the ability to plot detailed maps which clearly align with known chemical clustering. The metric space is given by compositional vectors in 103 dimensions and their relationships with respect to the EIMD.

This space and its induced structure has an associated geometry, but as we only possess the distances between points, we do not have the 2-dimensional coordinates that are required for plotting. We may use dimensionality reduction techniques to generate 2D Cartesian coordinates which respect the metric distances, and the resultant points are called an embedding of the space. By

measuring a line between embedded points with a ruler, distance may be used in the ordinary meaning of the word to define the similarity between two points. As it is generally impossible to represent a complex space without distorting the relationships between points, many dimensionality reduction techniques exist, each with their own focus at emphasizing specific relationships across a dataset.

In general we wish to align the distance between points with the associated EIMD between compositions, such that our embeddings give a valuable representation of the metric space. In this chapter we discuss embeddings produced by Uniform Manifold Approximation and Projection (UMAP)³², which gives clustered plots which allow the qualitative assessment of chemical datasets with unsupervised ML, and Principle Component Analysis (PCA) which we find gives more accurate representations of the relationship between points, with less overall distortion from the true positions.

In the UMAP algorithm, every composition is represented by a point and edges to each of the 15 most similar compounds in the dataset calculated with respect to the EIMD. It is not however possible to plot these distances directly, due to the contradictory information that arises when embedding a graph of high degree to the plane. Approximations of the metric distances are realized in two dimensional Euclidean space by constructing an inaccurate embedding of the point cloud, and refining the positions of the points along edges to each neighbour, such that distances between them align with the true EIMD, with respect to local cluster density. In doing this we disregard the majority of the inter-compound distances, yet retain a skeletal backbone which follows the local trends of the data, pulling together clusters of similar compounds. The resultant 2-dimensional plots are highly distorted from their true positions in the metric space, but in a manner which draws out the most prominent global trends of a dataset, from which we may pick out clear patterns and clusters both manually

and automatically. The exact positions of the points and associated clusters that make up the embedding are highly dependent on the choice of the hyperparameters, and particularly sensitive to choice of random seed. For the embeddings in this chapter we have chosen the default parameters (umap-learn version 0.3.6) and a `random_state` of 0, as these result in embeddings which have observable chemical trends. These plots are not perfectly reproducible due to the stochastic nature of the algorithm, however we have found that each random seed will lead to a similar placement and number of clusters in the resultant embedding.

For other datasets, a more directed approach may be taken by the trained chemist by taking a choice of random seeds and hyperparameters, and varying these until there is an embedding which is both visually pleasing and identifiably correct. The number of nearest neighbours each point should consider when refining edges in the lower dimensional space (default 15) is a common hyperparameter to vary. Changing the number of nearest neighbours to consider will alter the number of clusters in the embedding. If there is some prior knowledge of the number of clusters that should be contained in the dataset, this value may be modified until the embedding is in agreement with the known number of clusters. Additionally, the minimum distance (default 0.1) and spread (default 1.0) between embedded points may be adjusted to exaggerate how clustered/clumped each of the points in the embedding are. This may assist with visual clarity and benefit automated clustering techniques.

Binary Compositions

The binary compounds have simple compositions for us to demonstrate the efficacy of the EIMD and its alignment with domain knowledge. The 12,623 binary compounds in the ICSD were identified, the complete inter-compound distance matrix calculated with respect to the EIMD, and the resultant distances reduced to two dimensions with the application of UMAP using default parameters. UMAP is a stochastic embedding technique,

which can lead to entirely different embeddings depending on the choice of random seed and hyperparameters. Nevertheless, each of the clusters of clear separation tend to contain AB pairs from the same, or similar families on the periodic table, with trends across clusters following expected transitions in chemical composition through the modified Pettifor scale. In some clusters there is greater chemical discontinuity, yet there are trends across regions of these points which follow smooth variations in AB ratio. In others there is a strong chemical homogeneity, with A and B consisting of the same elements, with paths varying in x along the binary phase field of A_xB_{1-x} .

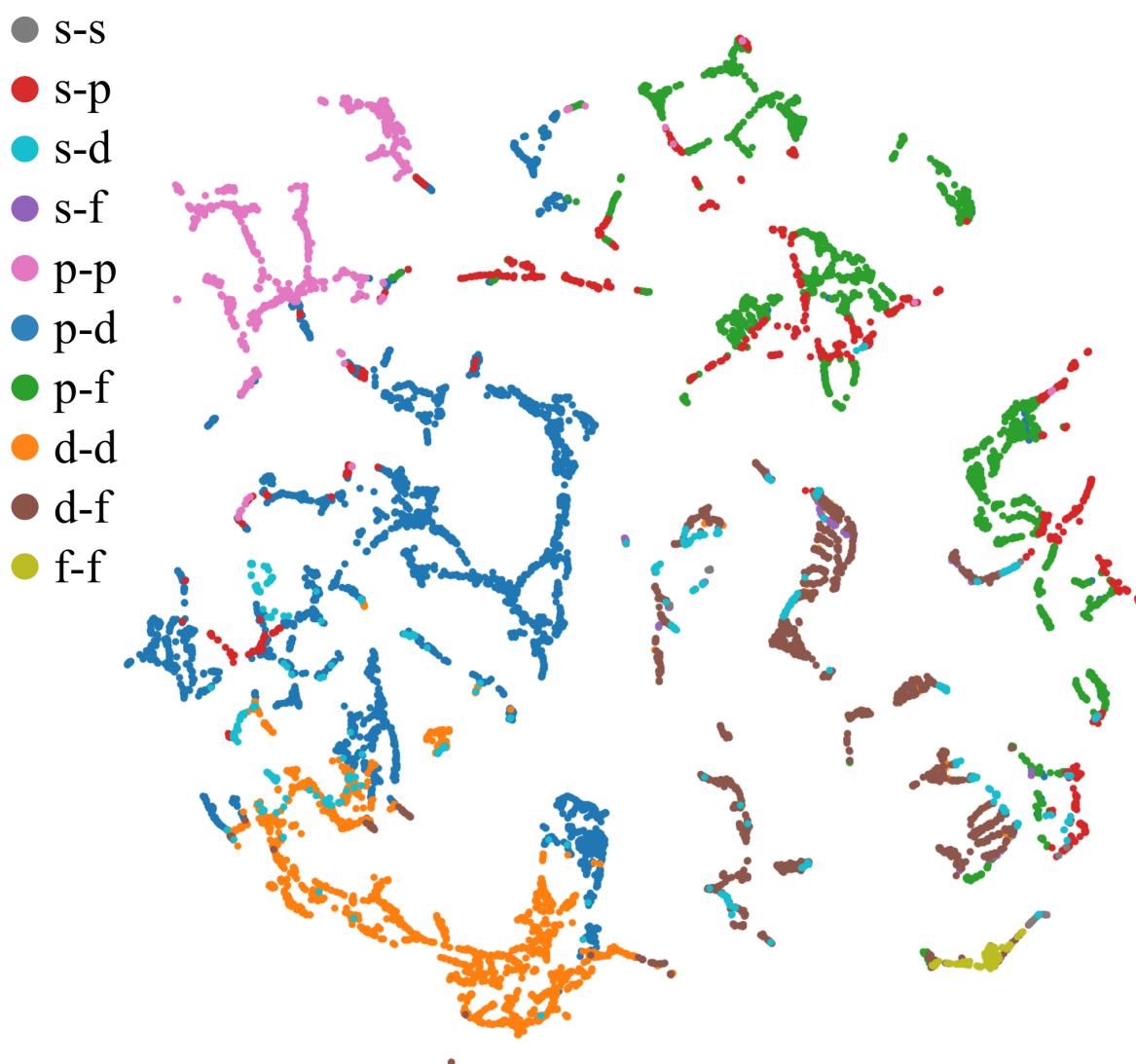


Figure 6: 12,623 binary composition vectors from the ICSD, with the EIMD between compounds calculated, and reduced to 2 dimensional coordinates using UMAP with its default parameters. With the application of colour labels to display the placement of elements present in the composition across the periodic table, it can be seen the EIMD has separated the space into complex-shaped clusters of related chemical families which have strong alignment with chemists' perception of similarity.

Adding chemical labels (*e.g.*, which blocks of the periodic table are found in the compound) demonstrate how trends in chemical properties are preserved when using the EIMD (Figure 6). As there are no experimental properties barring chemical formula and atomic positions reported in cifs, we must derive known features from the composition alone. In binary compositions, we know that the block of the periodic table that each of the two elements are from, will play a significant role in the resultant chemical properties. By labelling these blocks we can immediately see how UMAP has partitioned the space into clusters of compositions from the same, or arguably similar, blocks of the periodic table.

A clear example is the pink cluster in the upper left of the map, where we may find every compound in the ICSD containing two p-block elements. By embedding approximations of distances from the metric space, these chemical maps have been given a structure which aligns with domain knowledge. This alignment arises because the EIMD preserves chemical relationships between elements, and thus the chemical context present from the periodic table is present in the metric, which allows reference between elements (and by extension regions of the periodic table) ensuring these trends are well captured

Maps of inorganic compositional space have previously been created with the CED³³; however, for these maps to possess a structure which aligns with chemical judgement, any method employing CED requires a high incidence of compounds with shared elements. When similar methodology is applied to the entire periodic table *e.g.* binary compositions of the ICSD shown below, all detailed structure present when using the EIMD is lost.

When using the EIMD, compounds with elements from similar regions or blocks of the periodic table are clustered in groups with nontrivial shapes, high purity (*i.e.* a low number of labels per cluster), with a sensible relationship between clusters (Figure 6). When using the CED, all these desirable properties are lost (Figure 7); clusters have trivial shapes with little variation, are impure (*i.e.*

have a high number of labels in each cluster with combinations of each block in the periodic table evenly distributed across the map), and the clusters are evenly distributed throughout the projection. Further, rather than large, connected clusters generated when using ElMD, using the CED we get small and often isolated islands, with no clear relationship between these. Because compositions without shared elements have roughly equal distances under the CED, there are not enough global points of reference to place clusters in relation to one another with fidelity. There is dense clustering of compounds with similar stoichiometry (*i.e.* shared elements) due to the comparatively small distances between these, making it difficult to differentiate points within clusters.

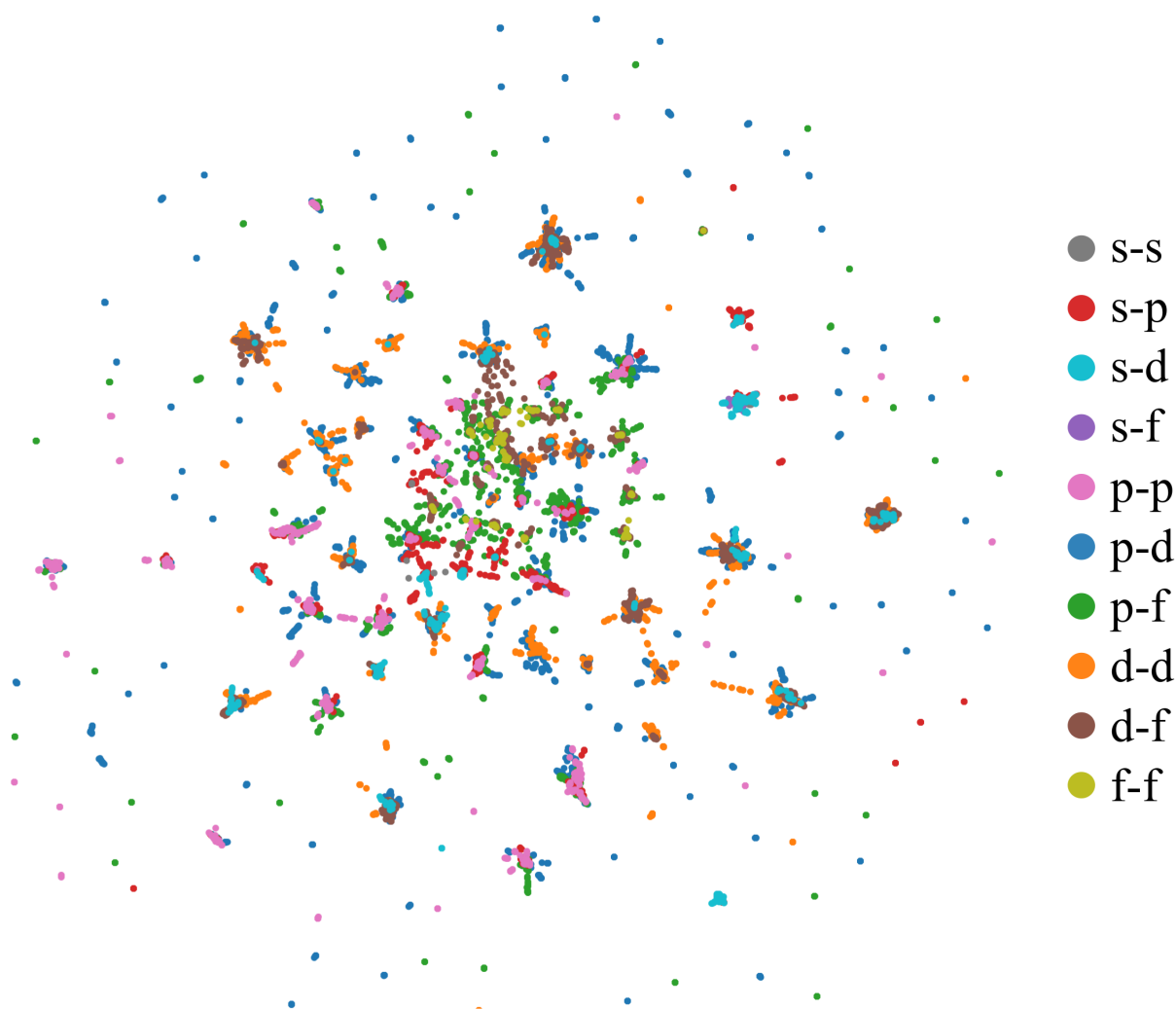


Figure 7: The maps produced with the CED contain many isolated clusters with trivial shapes and few members, with poor resolution of chemistry (*i.e.* many labels within a cluster).

This shows how a metric with qualitatively poor ability at distinguishing compositions will lead to quantitative confusion of known chemical relationships. Whilst the CED may provide enough distinguishing quality to be of benefit when applied to certain chemical datasets (e.g. where there are shared elements), the lack of chemical relationship between elements leads to a loss of discernibility and is a guaranteed source of noise for models. By processing compositional vectors with respect to the EIMD, chemical relationships and context have been preserved to a high enough standard that it may be captured reliably with automated methods.

The use of the EIMD enables the comparison between different choices of the elemental scale defining the indices in the compositional vector. This is not the case for the CED, where the distance is the same regardless of the elemental scale chosen. Even when using simple atomic numbers as the elemental index, the EIMD introduces significant structure to the UMAP generated clusters, leading to clusters with nontrivial shapes, however without the purity of labels observed when using the modified Pettifor scale. Elemental scales such as Pettifor's original Mendeleev number¹⁵ or alternate orderings of this scale³¹ result in plots with similar cluster shape and purity to the modified Pettifor scale (Supporting Information Figures S1-3).

An alternative approach to the use of compositional vectors X and Y is the use of recently developed vectors of features which are derived from values of physicochemical properties of the elements present in the composition³⁵⁻³⁷. Application of UMAP to the Euclidean distances between the magpie features³⁸ of these binary compounds results in clusters with low levels of chemical purity, similar to the results obtained using the EIMD with atomic number scale (Supporting Information Figure S1).

The labels in Figure 4a were assigned with the density based clustering algorithm DBSCAN³⁹ on the points obtained by the EIMD and UMAP, which assigns groups of points class labels,

such that clusters which have been closely plotted together on the plane share a label. As there are few pre-existing chemical properties we can attribute to this dataset, we rely on unsupervised learning to gain insights from our data with ML.

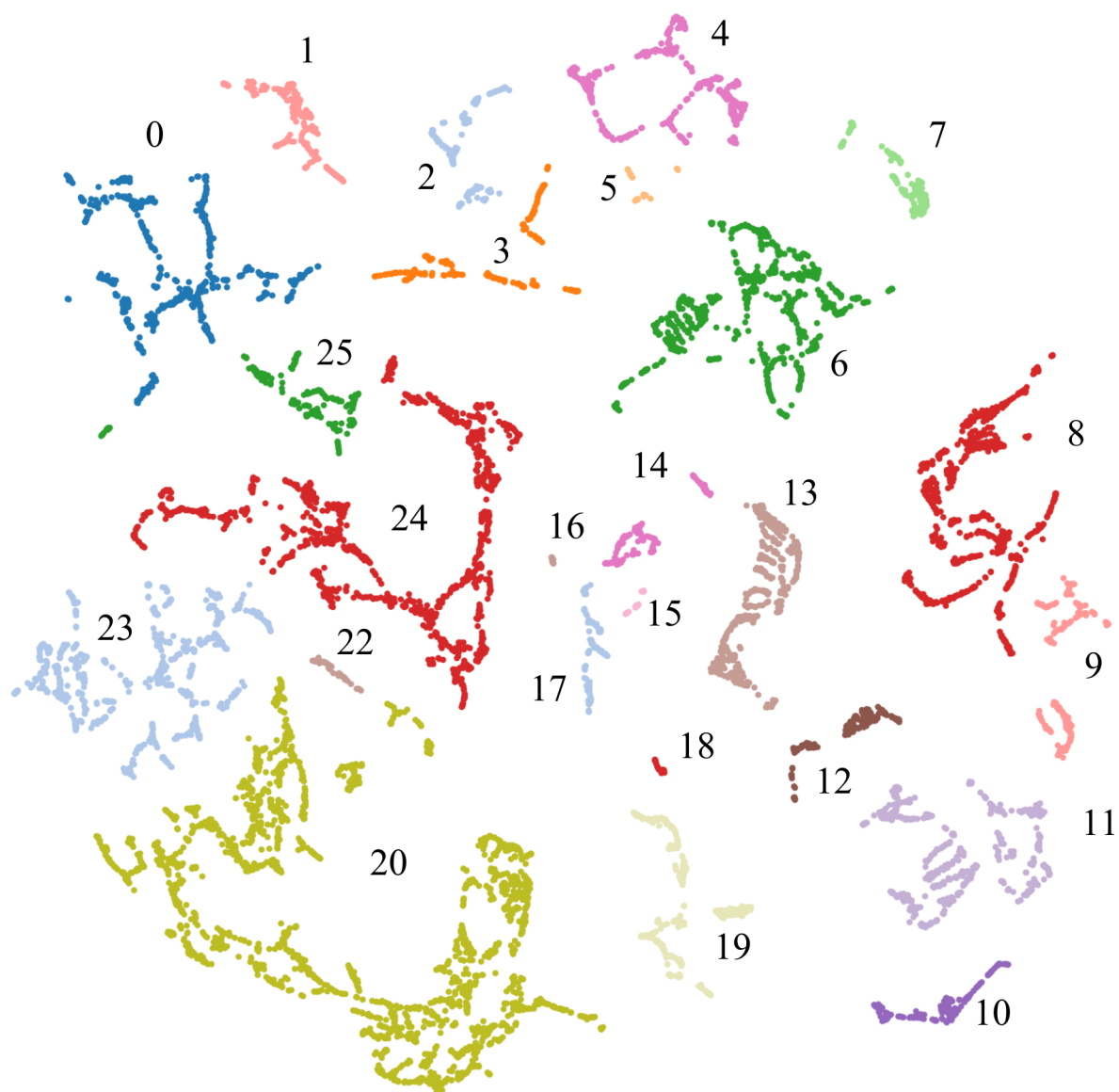
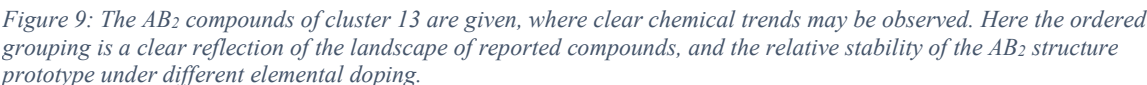


Figure 8: The same embedding of binary compounds from Figure 6, segmented into 26 distinct clusters using the DBSCAN algorithm. A complete analysis is provided in the appendix of this chapter.



After cluster 13, the second clearest example of parallel trendlines in chemical features can be seen to the left of cluster 6, with the general form AB_3 , A being an f-block metal, and B being a p-block metal. From the top left of the cluster to the bottom right, A ions follow the Pettifor scale. Across each successive line from left to right, the B ions progress through Al, Ga, In, Tl, Pb, Sn, and finally Ge. Complete analysis of each DBSCAN cluster is given in Supplementary Note 1. With no prior chemical knowledge of these compounds, we can draw attention to underlying chemical properties, providing visually qualitative maps, capturing families of clear relation.

The Inorganic Crystal Structure Database

For each of the 125,627 unique compound formula in the ICSD we may apply the same process, but due to the scale of the task defining clusters becomes difficult. We may instead attribute known chemistry about each composition to uncover underlying trends in the data. In Figure 5a we see these compounds plotted via UMAP, and coloured by taking the standard deviation of the respective electronegativities of the constituent atoms. We calculate this by taking the associated Pauling electronegativity for each of the non-zero elements in a compositional vector, giving a set of electronegativities, E , of length n . The average electronegativity of the set, \bar{e} is calculated, and standard deviation obtained via the standard formula:

$$SD = \sqrt{\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n - 1}}$$

This simple measure reveals a clear trend in chemical property across the reported compounds, between the more ionic compounds across the upper right side of Figure 5a, to the more covalently bonded across the left boundary.

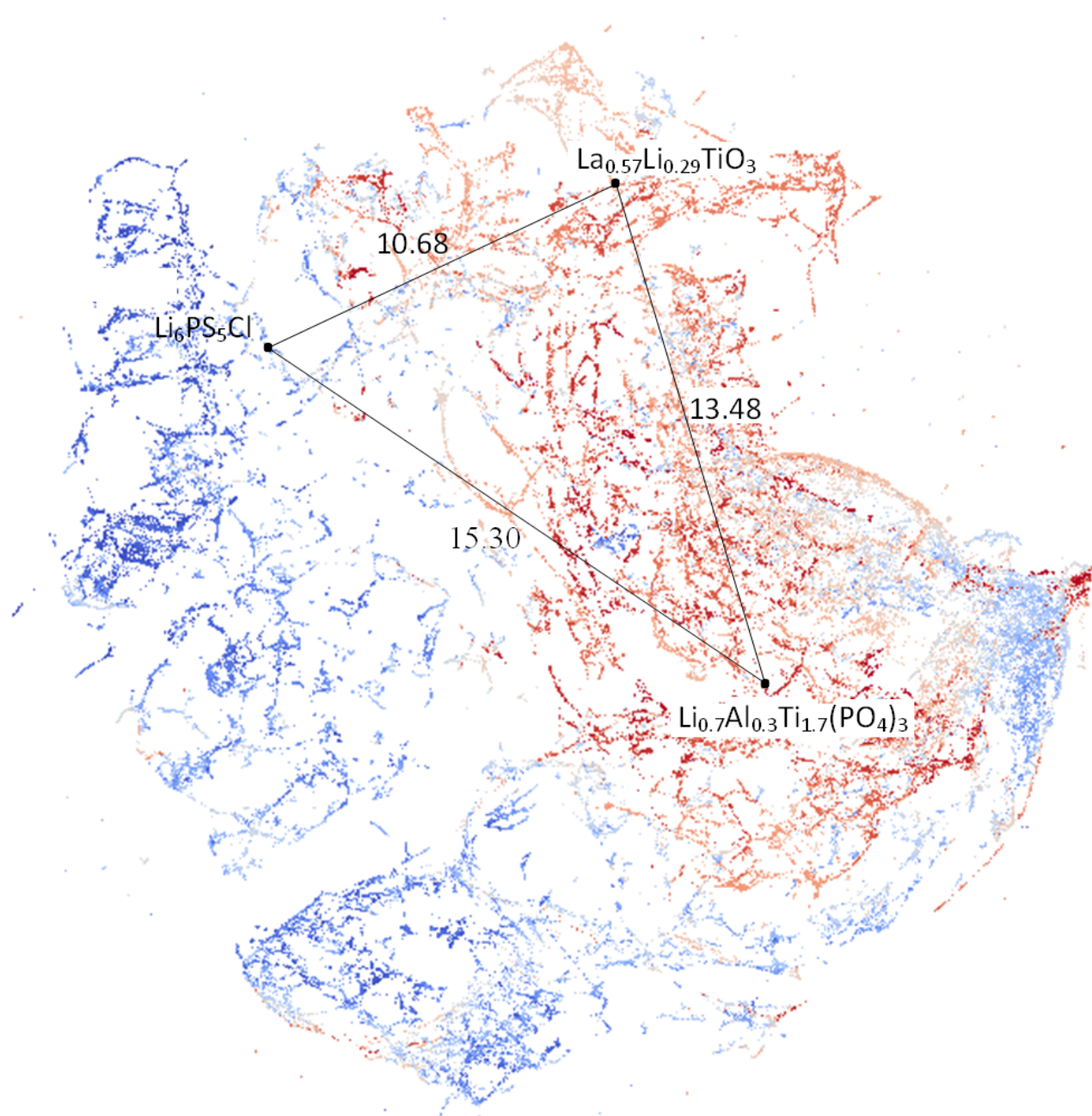


Figure 10: 125,627 compositions from the ICSD with their inter-compound ElMD calculated, and resultant distances reduced to 2 dimensional coordinates with UMAP. Three candidate solid state electrolytes are overlaid with the planar distances between points labelled. Standard deviation of electronegativity for elements in each compound is given by the colouring from red (more covalent) to blue (more ionic). It can be seen that UMAP has accentuated some of the more subtle aspects of chemical similarity by distorting the space.

It should be noted that the UMAP algorithm emphasizes the clusters of a metric space. When optimizing distances, UMAP ensures that clusters of compounds are closely packed within families and clearly separated from other clusters on the plane. Unsupervised density-based clustering algorithms such as DBSCAN therefore work consistently and effectively on the resultant plots, allowing the swift classification of new compounds from existing knowledge.

Whilst local neighbourhoods will have similar structure to the metric space, the global trends appear warped. This is highlighted by the three SSEs from the previous section overlaid on Figure 5. It can be seen that these do follow the approximate similarity given to us from EIMD, but have been distorted from the perfect line they fall on in the EIMD metric space. We may take the local distance between each of the embedded points, and by calculating the Pearson's correlation between each of these and their associated EIMD, the quality of these embeddings may be assessed.

While UMAP has given value by separating these projections into clusters of familial relation, by referring to Table 2, with a correlation of 0.748 many of the distances have been distorted from their true values, making these potentially unsuitable input for regression tasks. We would expect that reducing our distances to higher dimensional coordinate systems would give UMAP more degrees of freedom when embedding a graph layout, however past two dimensions the correlation does not improve, in part due to the implementations focus on planar projections.

Principle Component Analysis

A truer picture of the metric space may be obtained via principle component analysis (PCA), a widely used dimensionality reduction technique in the natural sciences for projecting data along axes of greatest variance. We will briefly outline the modifications that

must be taken from the standard procedure to allow the inspection of the ICSD using PCA with respect to the EIMD.

In the normal PCA process, one takes the high dimensional data vectors and computes the covariance matrix, C , between each datapoint. The eigenvectors of C are computed via singular value decomposition (SVD) then sorted by their largest eigenvalues, with the two largest eigenvectors selected as the new co-ordinate axes to project each data point to.

In this instance, we do not wish to use the composition vector covariance matrix, as this will correspond to the CED eigenvectors. Instead, we want to use the EIMD kernel matrix to construct our embedding, so we may observe the trends in the EIMD function space. Thankfully, there is a technique which allows non-linear distance functions, such as the EIMD, to be embedded to 2-dimensional Euclidean space, called kernel PCA ⁴⁰, which follows a nearly identical process to traditional PCA.

An EIMD distance matrix is considered as a kernel matrix of the metric function space. If the chosen similarity function is not a metric, this kernel may be squared to remove negative values. Next, the kernel matrix goes through a process known as double centering, which involves subtracting the row and column means to give a new axes mean of zero, and then scaling this by dividing by the number of rows/columns. The final matrix is called a Gram, or Gramian, matrix. Applying SVD to the Gramian and sorting by eigenvalues allows us to select the geodesic contour eigenfunctions through the function space which show greatest variance. Projecting the points obtained through SVD along these axes has the effect of “flattening” the curvature of the space out, but preserves a scaling of the distances in the metric space ⁴¹.

In practice PCA will compress each of the distances linearly. As there is rarely an embedding of a high dimensional pointcloud in lower dimensional Euclidean space which will respect the global structure perfectly, this often creates overcrowded plots with a loss of intrinsic structure. When applied to our dataset this does create

overly dense regions of points, making these embeddings unsuitable for the automated identification of clusters with an algorithm such as DBSCAN. These embeddings still retain a demonstrably strong resemblance to the true structure of the metric space however, which is presented in Figure 11.

We have found that even in lower dimensional spaces, the local Euclidean distances between points retains a reasonably high degree of correlation with the EIMD (Table 2). In 3 dimensions, with a correlation of 0.945, we may take these as semi-reliable reduced composition vectors with respect to the embedded 3-dimensional Euclidean distance. Embedding to higher dimensions with PCA does not improve on this correlation as the underlying space is seen to be approximately 2-dimensional ⁴² with an observed global saddle shape in 3 dimensions.

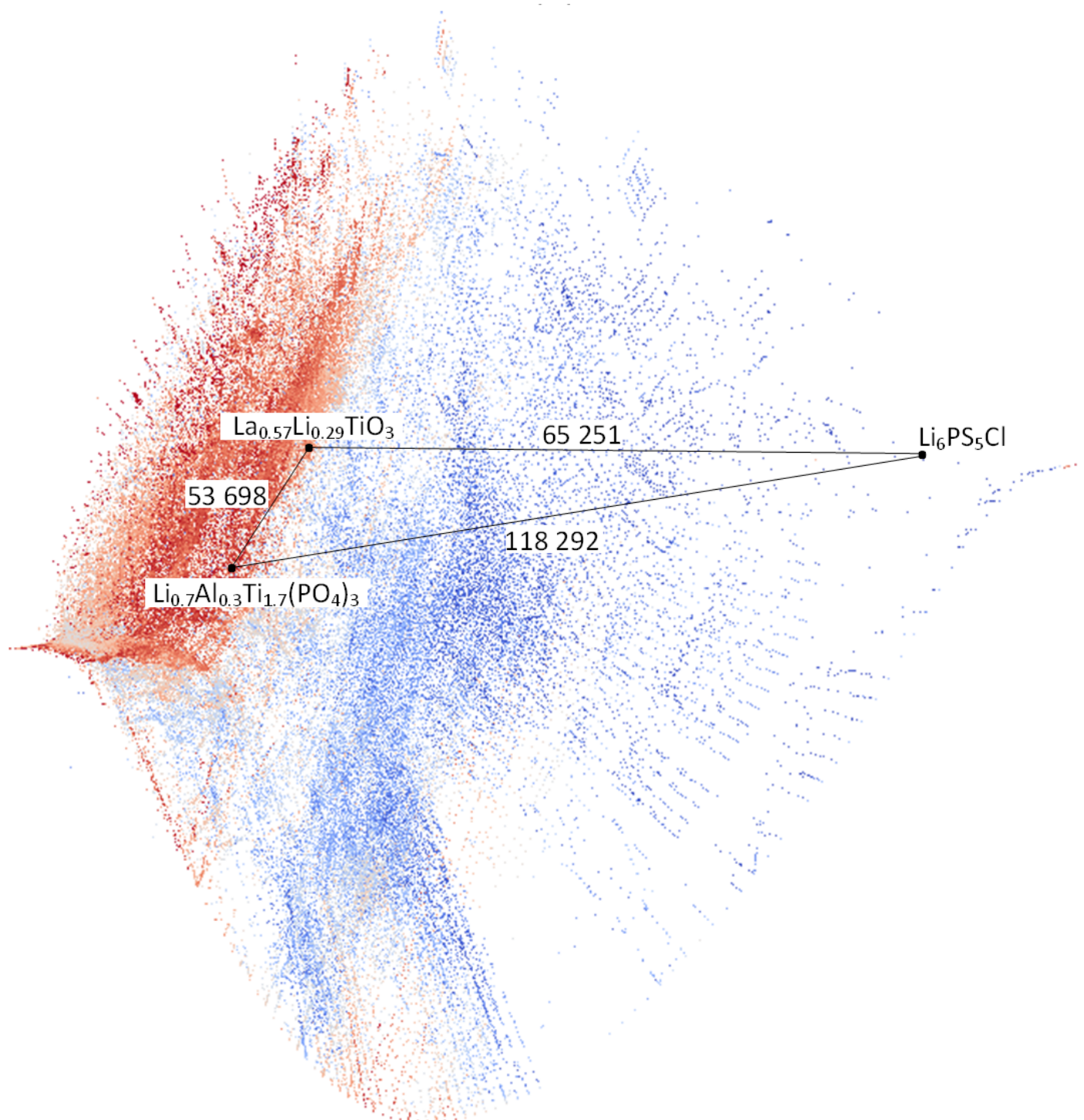


Figure 11: 125,627 compositions from the ICSD with their inter-compound EIMD calculated, and resultant distances reduced to 2 dimensional coordinates with PCA. Three candidate solid state electrolytes are overlaid with the planar distances between points labelled. Standard deviation of electronegativity for elements in each compound is given by the colouring from red (more covalent) to blue (more ionic).

Table 2. Pearsons correlation coefficient between the complete Euclidean distance matrix for embedded points, and true ElMD distances between compounds in successively higher dimensional embeddings

| Embedded Dimension | UMAP | PCA |
|--------------------|-------|-------|
| 1 | 0.538 | 0.860 |
| 2 | 0.748 | 0.938 |
| 3 | 0.736 | 0.945 |
| 5 | 0.661 | 0.945 |

It can be seen that each of the single elements may be found along the bottom edge of the plot, at the tip of each of the parabolas. Trending away from these are the associated binary and ternary compounds in divergent lines of placement. We can clearly see the abundance and scarcity of reported compositions containing certain elements along the modified Pettifor scale, and trends in chemical makeup can be observed. Whilst this may not give us the best map of compositions for effective ML, it remains valuable for its accurate realization of the metric space. This enables us to map the chemical relationships between all of the compositions in the ICSD, with confidence that our embedding is representative of the relation between compounds given to us with the metric, which may be explored interactively at www.elmd.io/plots/.

The Pearsons Database

The Pearsons database, based on the Pauling File ⁴³, consists of 354,573 structures with an interactive GUI. At time of writing there is no method of accessing these structure files in a high throughput manner, however each of the compositions may be retrieved in csv format through the interface, each of which have a label with the associated structural prototype of the crystal. Whilst 11,688 of the materials have not been assigned a structural prototype label, the remaining 342,885 entries are labelled, which means there are of the 181,786 unique compositions , there are 174,866 with an associated prototype structural label. The inter-

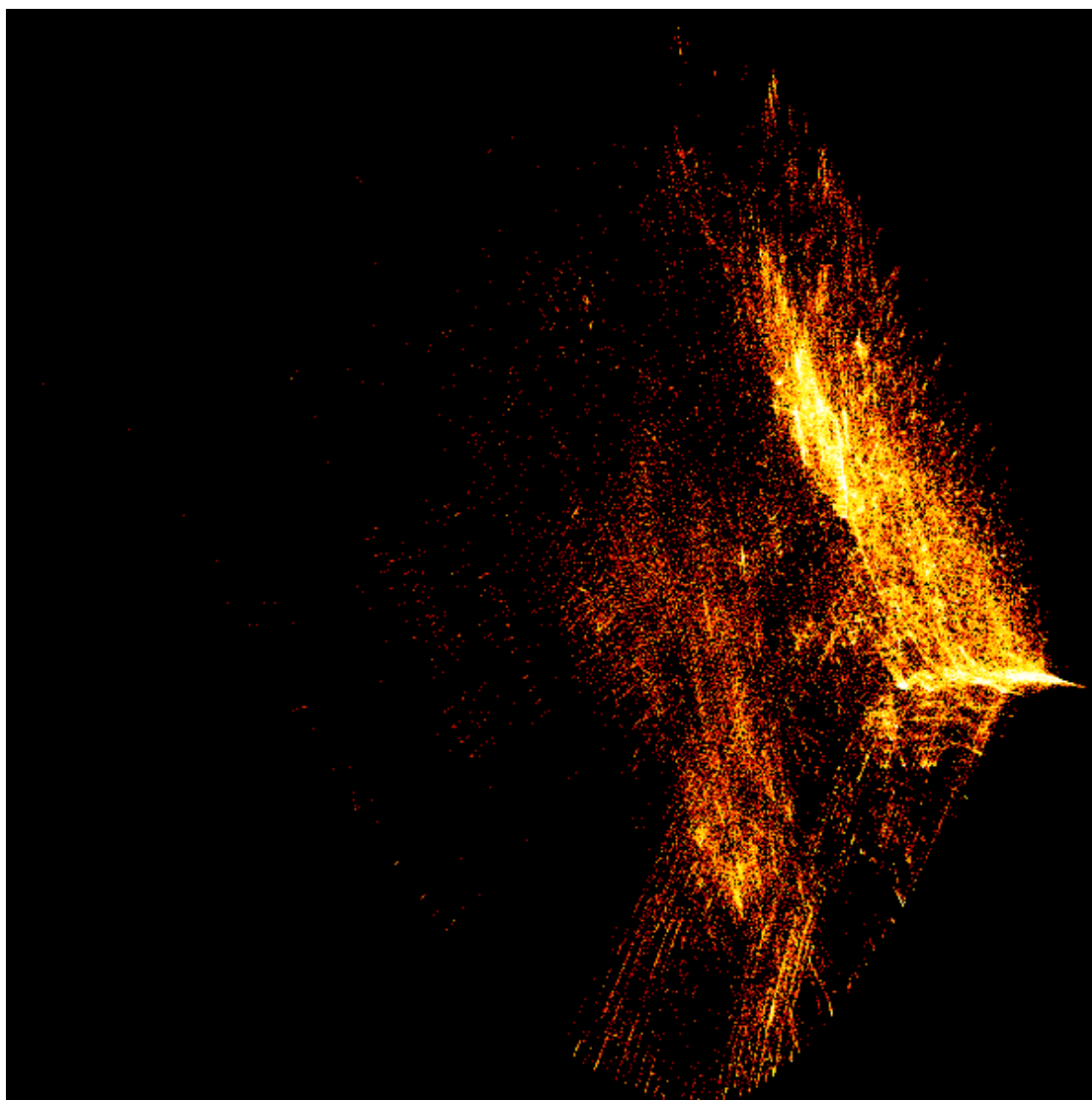


Figure 12: 181,786 unique compositions from the Pearsons database embed with iterative PCA with respect to the EIMD between compounds. The yellow intensity has been shaded according to the relative density of points which fall under each pixel in the image, allowing us to observe global chemical trends in the materials deposited in this dataset.

composition EIMD kernel matrix is formed and converted to its centered Gramian. Given the expanded size of the dataset, with no modifications, the standard kernel PCA process described earlier requires in excess of 1TB of computer memory, which makes it challenging to operate with current hardware. These distances are instead embedded to the plane using incremental PCA⁴⁴, which takes the first 20,000 datapoints, and computes the principle components of these. Successive batches of size 20,000 are projected to these principle components, which are then adjusted to account for the new variance introduced by each batch. Once the final singular vectors are computed, the initial data is projected to these new axes, with the *x*-axis inverted so the plot is aligned with the modified Pettifor scale. Due to the large number of points, a density plot of the embedded points gives a more representative overview of the space (Figure 12).

The first observation is that the Pearsons PCA embedding is visually similar to the ICSD embedding. Regions of compositional stability are known, and have been well studied and reported. This is reflected by the dense regions that are shared across the two plots in the left, upper left, and lower regions of each map. This is also unsurprising, as there are 86,194 shared compositions between the two datasets, with Pearsons being the larger superset of the two.

We may use the elemental information and the structural prototype labels to reduce the observable chemical information to make sense of this space. There are 36,710 unique structural labels contained in the Pearsons database, which is too many to be of practical use in a labelled plot. Reducing the chemical information gives us general points of reference. Focussing on some of the most common structural families reveals that with no structural information present, structurally similar compounds are located in similar regions of the embedding. Simpler prototype structures are found in the lower region of the map, with more complex structure types encountered as we move towards the upper right region.

By first isolating each of the single element compounds we see how these distribute themselves along the lower boundary of the embedding by their assigned modified Pettifor number. Taking the distribution of elements along this scale, we see how each pure element is represented by a single bar which follows the Pettifor number as we move along the plot (Figure 13). This highlights the relationships already encoded by the modified Pettifor number, but not how the space considers their combinations.

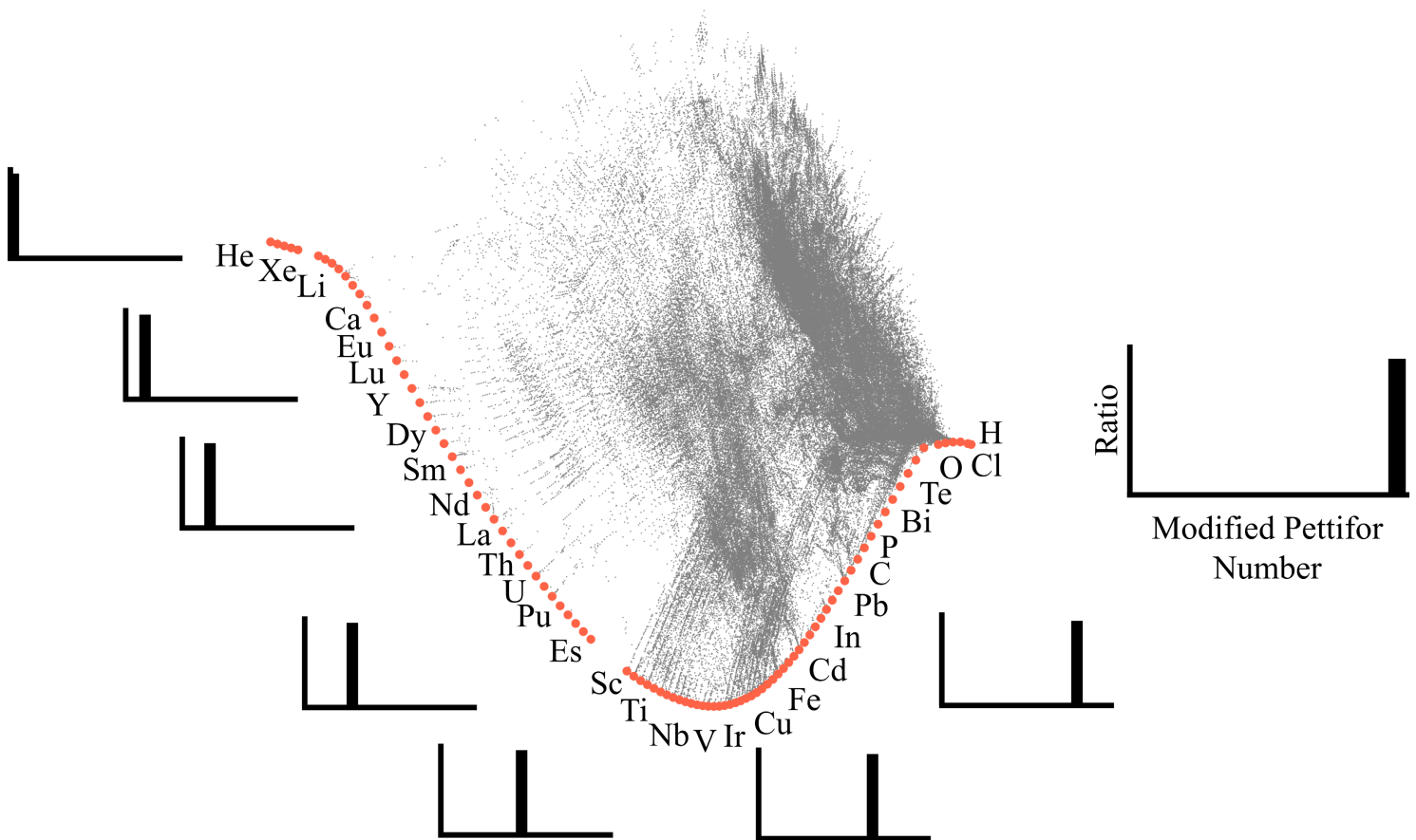


Figure 13: The pure element materials may be found along the lower boundary of the map in increasing assignment of modified Pettifor number. The general trend charts the noble gases, followed by the s-block, f-block, d-block, p-block, then hydrogen.

The simplest method of placing points of reference is to place each of the AB binary compositions over the projection and mark the changes in elemental distribution as we process around the plot. The binary compounds with similar AB elements are found along the lower boundary, as before. Going from the lower centre of the map upwards shows increasing divergence in modified Pettifor number until the s-block halides are encountered at the top of the map. Following the embedding along either of the upper regions returns compositions with a low modified Pettifor number with the partners modified Pettifor numbered lowering as we progress down to the left, and vice versa to the right.

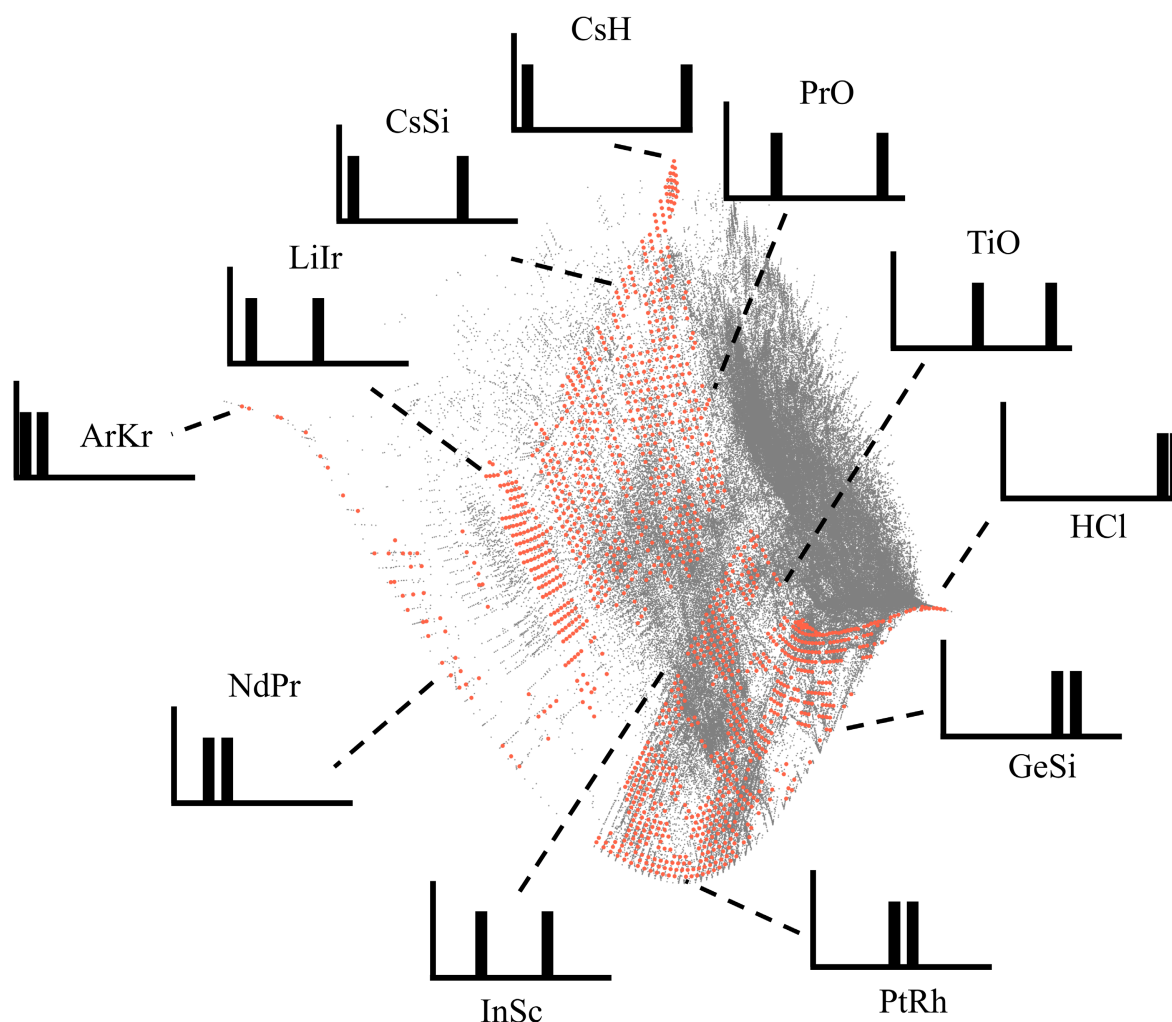


Figure 14: The AB binary compounds reported in Pearsons follow geodesic paths in ElMD space, with piecewise linear step changes occurring with each feasible substitution. AB compounds with highly similar elements follow the lower boundary in alignment with the modified Pettifor number, as before. Going from the lower tip upwards, we see distributions which are symmetric around the central Pettifor mass with increasingly divergent elements. Mass is increasingly weighted towards each of the poles as we follow the path from the utmost point to the left and to the right.

Adding the remaining binary compounds with the general form A_xB_{1-x} to the embedding creates a denser plot. Whilst specific patterns are harder to make out, it is clear that the underlying geodesic paths that could be observed with the AB compounds remain present. This is unsurprising, as these remaining compositions will linearly interpolate the space between each of the elements and their associated AB compounds. The upper left and right regions are populated with A_xB_{1-x} compounds where the ratio of A is larger than B in the upper left, and vice versa for the upper right.

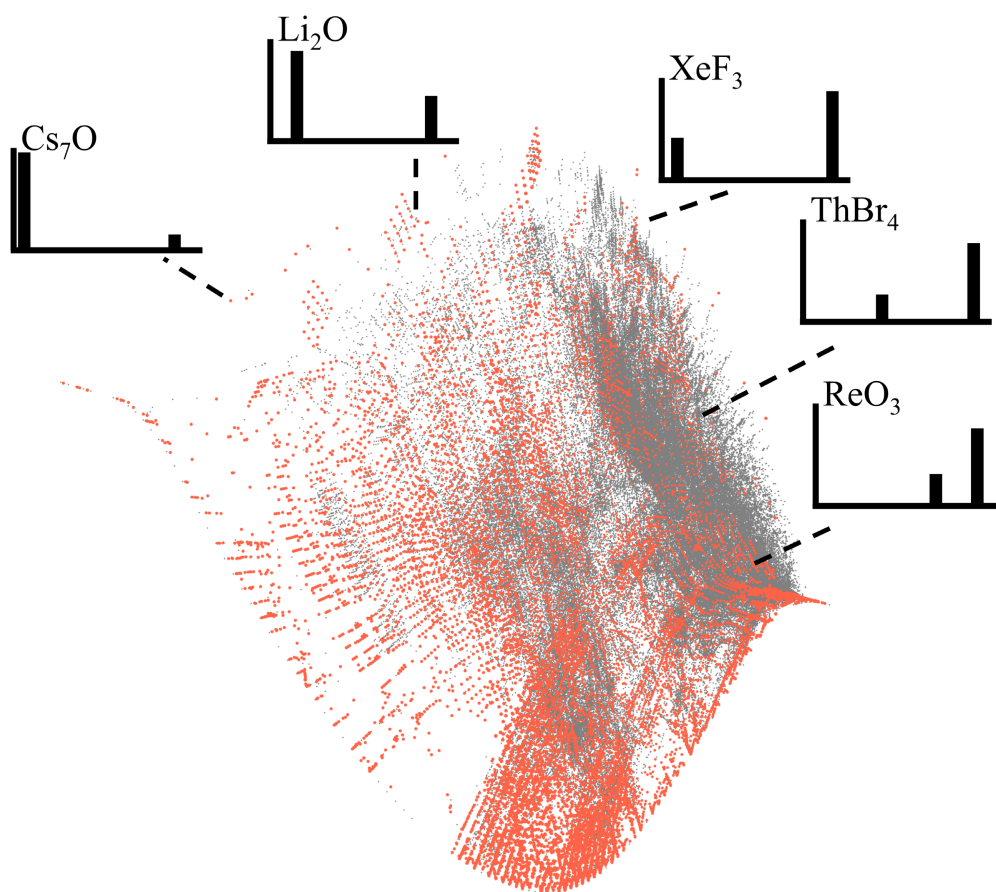


Figure 15: A_xB compounds are distributed along similar paths to the AB compounds, with a greater population of materials up and to the left where the molar ratio of $A \gg B$, and $B \gg A$.

Geodesic paths of materials with the A_xB_{1-x} compositional type can be traced across the embedding. When these compositions are extracted it is found that these either follow trends in substitutions, or trace a value of x across a series of materials with the general form A_xB_{1-x} . Trends in x doping can be clearly seen in the series running from V to O, with O doping increasing from $\text{VO}_{0.03}$ to V_2O_5 .

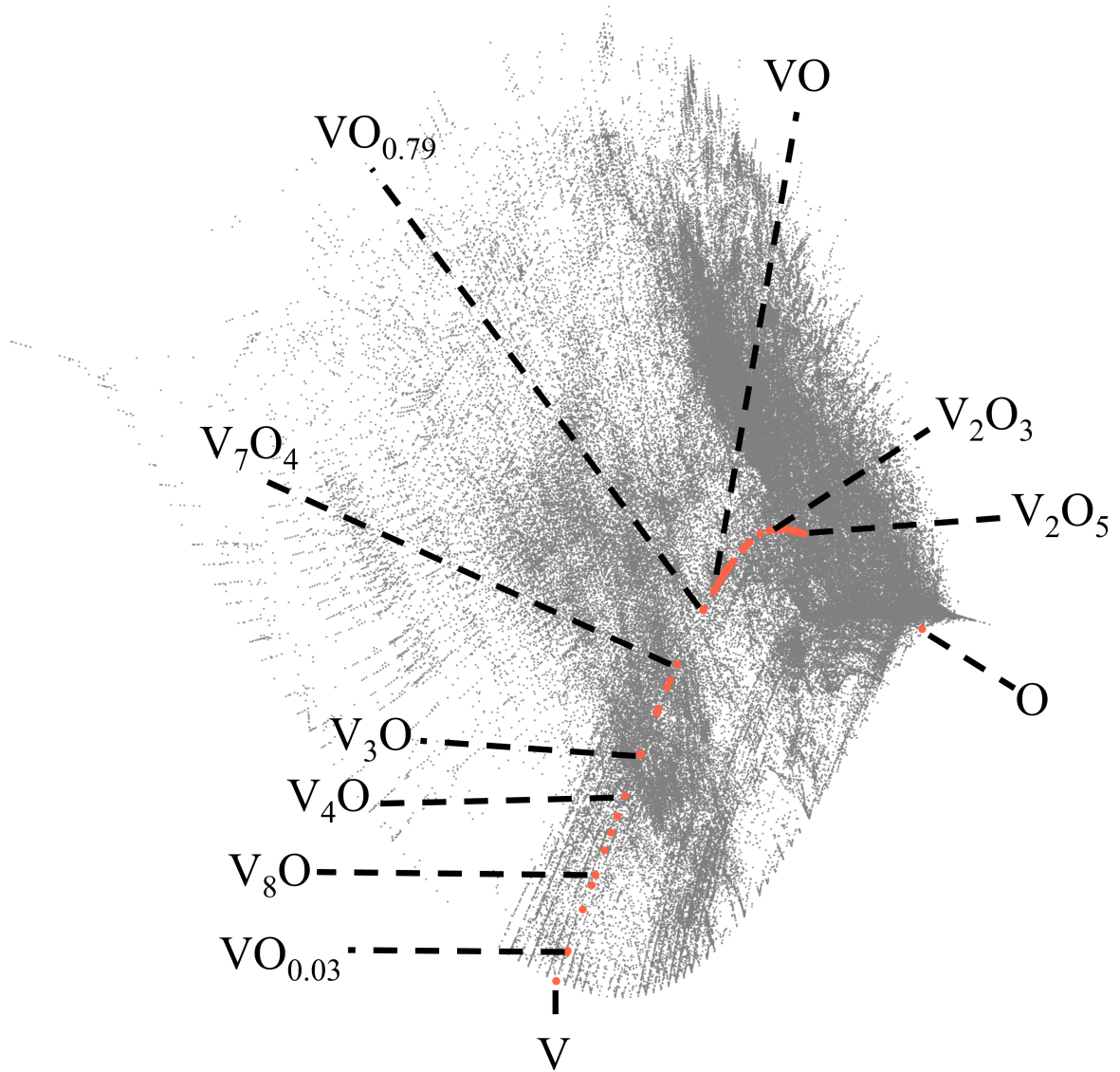


Figure 16: A path of known stable vanadium oxides traced across the embedding. In the function space, these can be plotted on a straight line geodesic. This betrays the curvature of the space that is lost when embedding in 2-dimensions.

Examples of sequences of known feasible substitutions are found throughout the map, and can be seen in the transition metal intermetallic alloys of AB compounds, where the A and B ions increase or decrease in Pettifor number as we follow paths towards the respective elements through the space.

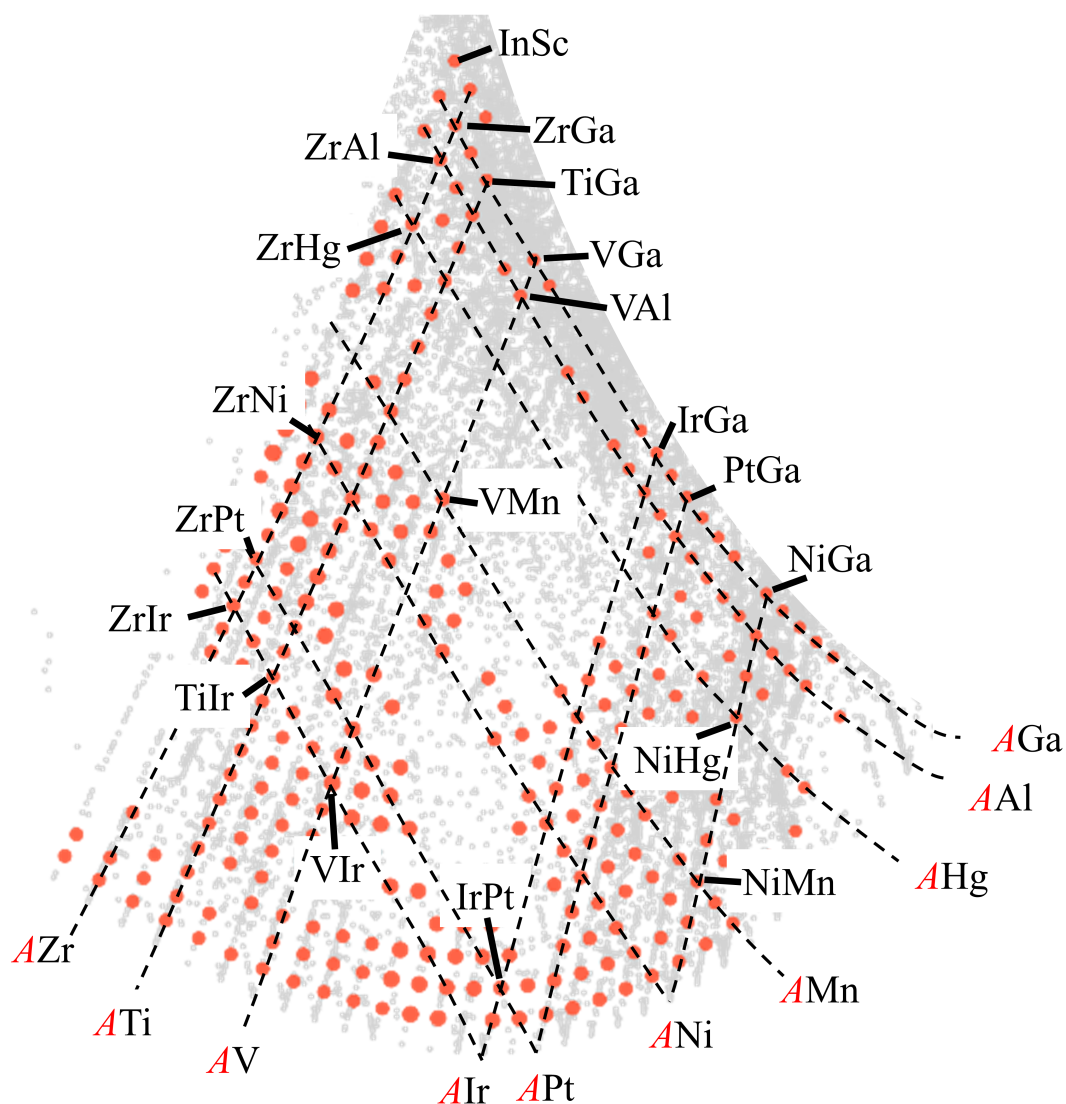


Figure 17: A zoom of the transition metal region of the EIMD embedding, with reported AB intermetallic compounds in the Pearson's dataset overlaid in red.

A similar filtering operation may be applied to the dataset to highlight all compositions which have the potential to be perovskite compounds. Perovskites tend to have the general form ABC_3 and form in a cubic arrangement, which may be filtered compositionally by highlighting all materials with a $\{0.2, 0.2, 0.6\}$ compositional ratio. In doing so, we see how chemically similar materials cluster amongst one another in an ordered fashion. The general distributions of each of these clusters are observed to follow the same general trends as the binary compounds by the distribution of mass along the modified Pettifor scale. These observations allow give us some intuition to the general ways that elements are arranged when navigating the space.

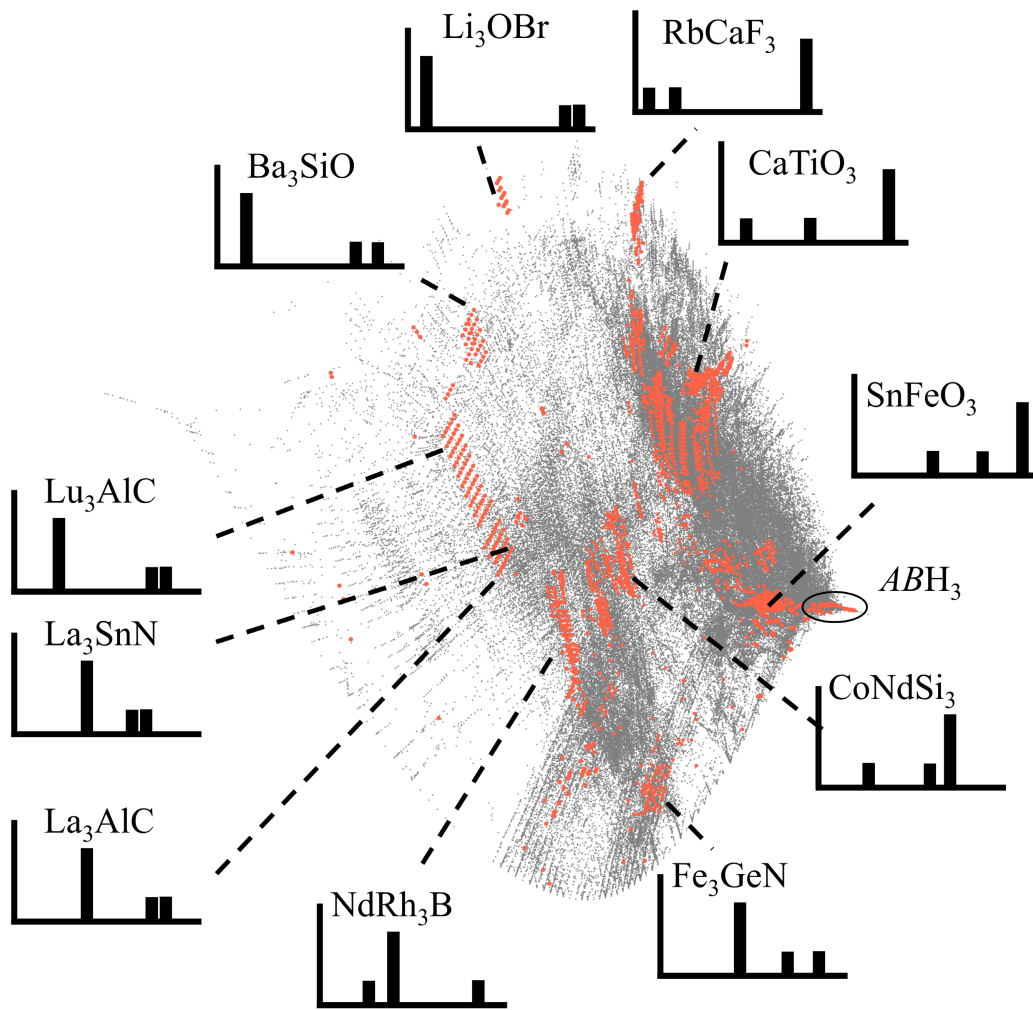


Figure 18: The ABC_3 compositions of the Pearson's dataset overlaid on the EIMD embedding. The distribution of elements follows similar trends to the AB compounds in their placement of mass along the modified Pettifor scale.

Applying the structural labels provided by the Pearsons dataset, it may be seen that there are 3,882 compositions which are labelled with one of the 61 single element packings contained in the dataset. By plotting each of these on the embedding, it is seen that the vast majority of these structures fall along the lower boundary of the map. There are too many unique structural labels to meaningfully digest the information, and as such we shall isolate some of the more common labels in turn.

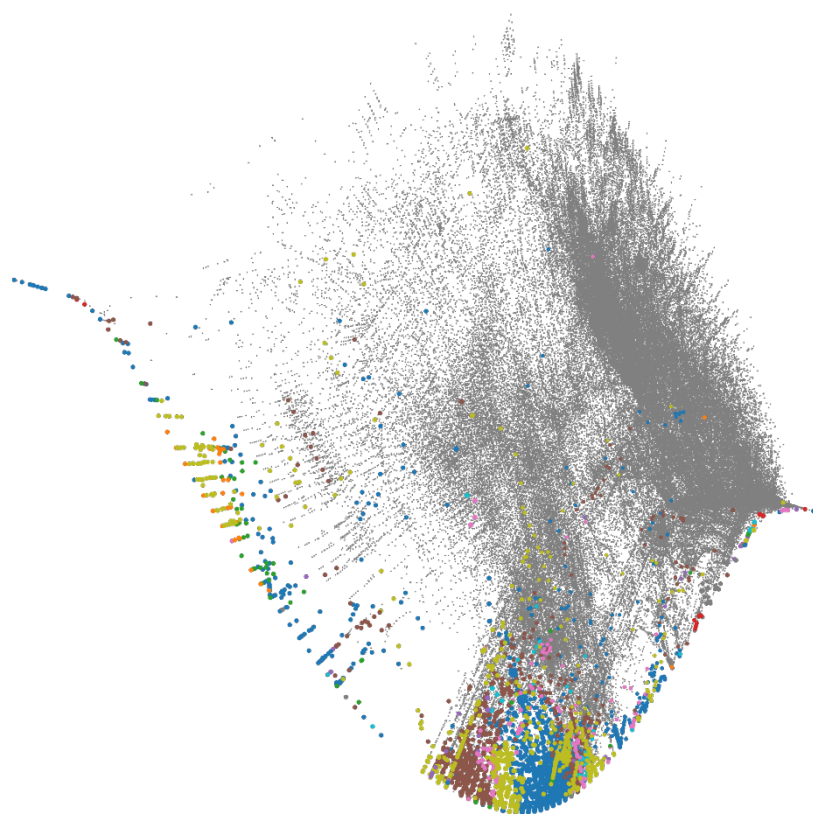


Figure 19: In the pearsons dataset, 3,882 structures have one of 61 single element structural prototypes

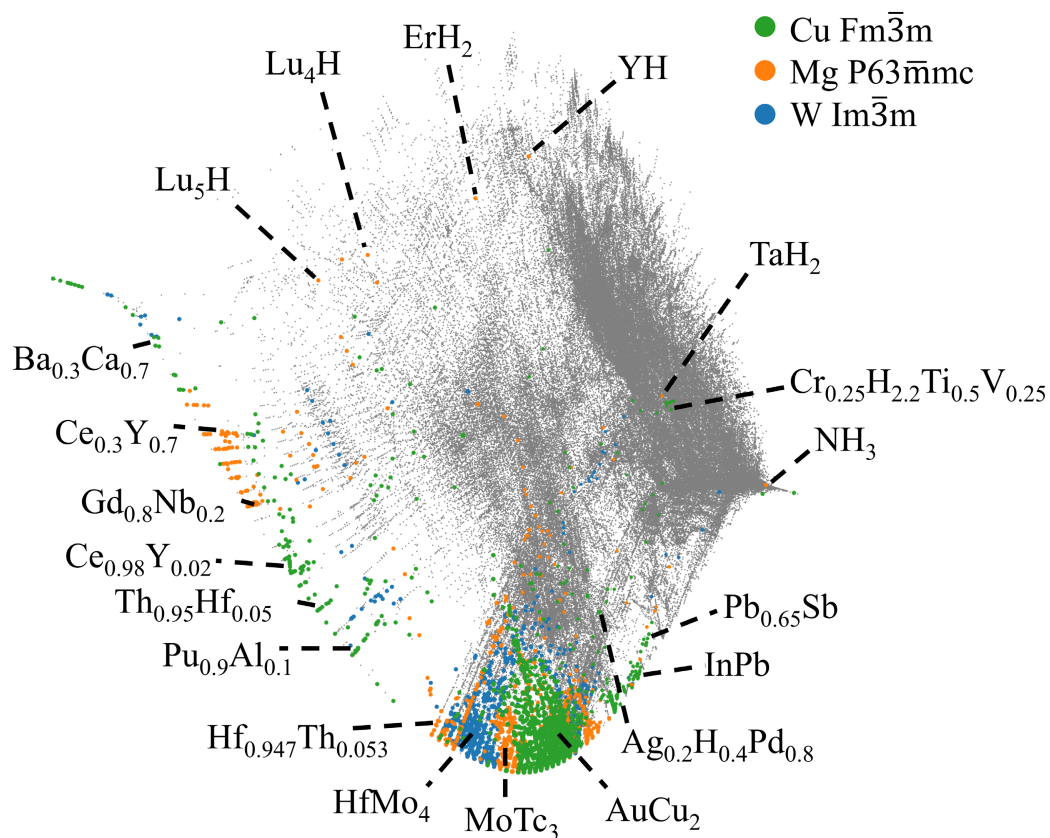


Figure 20: The three most common single element structural prototypes contained in the dataset.

The three most populous single element structure prototypes, the Cu $Fm\bar{3}m$ (1,632 compositions), Mg $P63\bar{m}mc$ (712 compositions), and W $Im\bar{3}m$ (932 compositions) structure prototypes exemplify this (Figure 19). These compounds lie along the bottom boundary of the map before diverging into a wider spectrum of structure types as dopants and stoichiometries are varied moving up the plot.

The cubic binary structure prototype, NaCl $Fm\bar{3}m$, shown in Figure 21, is a stable packing and many (2,648 compositions) structures take this arrangement. Highly imbalanced metallic binary compounds with rock salt phases, such as PtB_{0.01}, are found at the bottom of the map. Moving up the embedding gives greater diversity in the ratios of the intermetallic elements, with the number of dopants and the divergence in modified Pettifor number between the elements of the composition, increasing as we go up the plot. With this label, selenides, tellurides and oxides with p-block and more electronegative transition metal dopants are found to the right of the embedding, and carbides and nitrides found in the centre. Lanthanide and actinide containing compounds are found in the central upper region of the plot. The isolated cluster at the very top contains halide binary compositions whose elements possess highly dissimilar electronegativities, including the chloride salt NaCl.

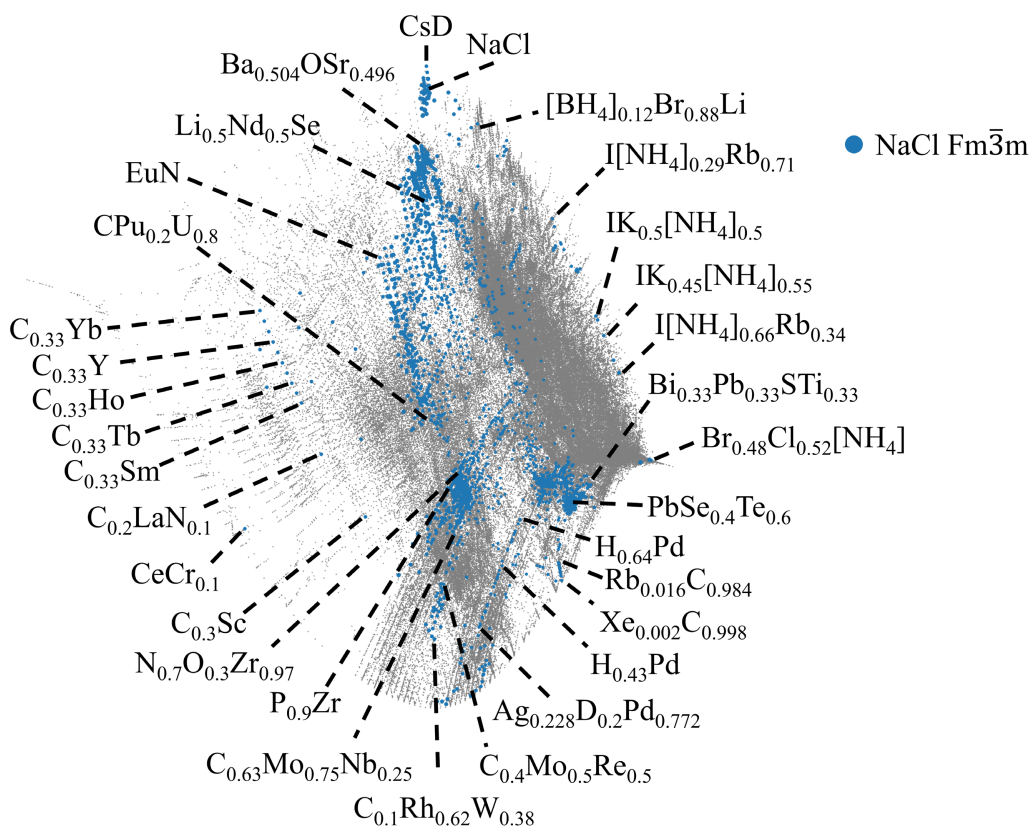


Figure 21: 2,648 compounds in the Pearson's dataset with the NaCl rock salt structural prototype. The compound NaCl itself may be found in the upmost cluster of the embedding.

In Figure 22, compounds with a structure prototype similar to ZnS $F\bar{4}3m$ (677 compositions) can be found between the NaCl selenide/tellurides and the NaCl oxides. The $Pm\bar{3}m$ CsCl structure prototype (968 compositions) covers many of the less electronegative transition metal intermetallic structures towards the bottom of the plot. These CsCl structures also expose trends in the f-block intermetallic structures, and continue trends that are present in the NaCl compositions. Changing the kind of material under consideration, let us now examine structure types with the general formula AB_2 .

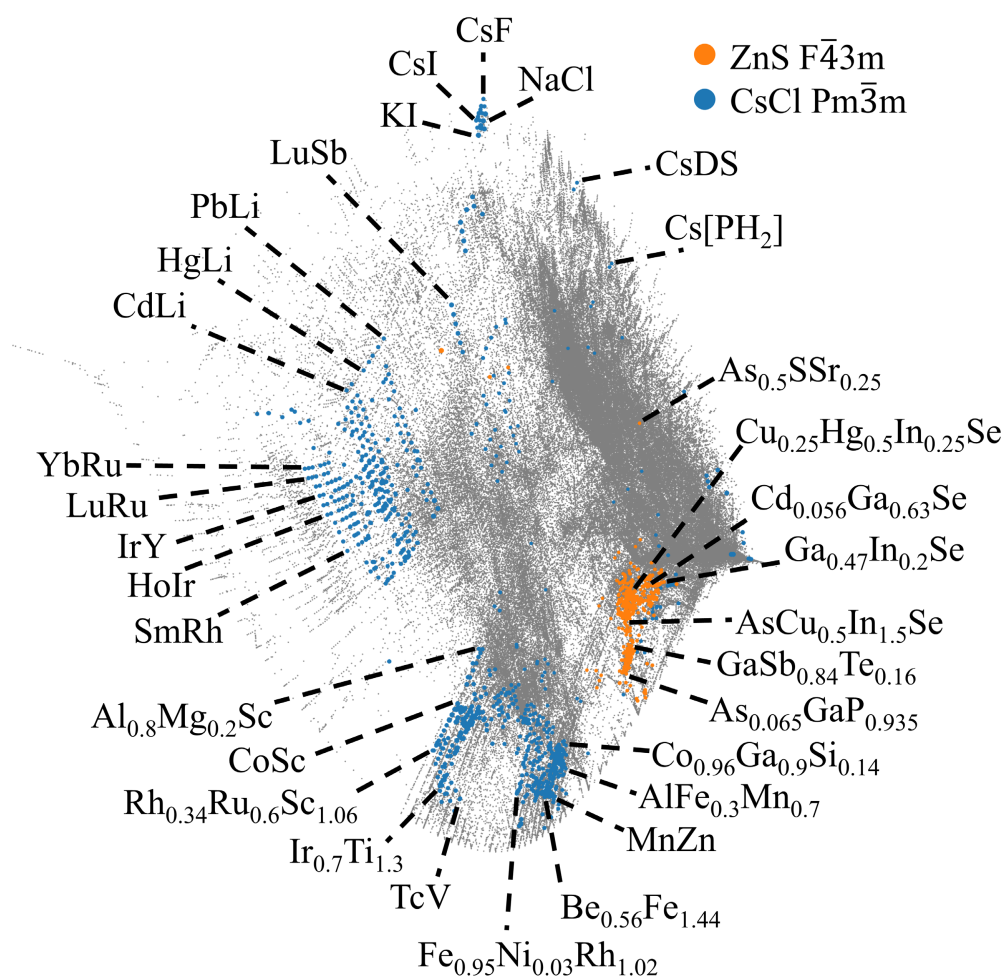


Figure 22: ZnS and CsCl are two commonly reported structural prototypes. These compounds are encountered in similar regions to the NaCl structure type, barring the leftmost region, which continues trends present in the NaCl substitution paths.

The most populous of which, CaF_2 $\text{Fm}\bar{3}\text{m}$ with 2,357 compositions, is found in the upper right region of the map shown in Figure 23, with reasonable separation from the outer hull of the embedding. Compounds found in this region are metallic halides, metallic hydrides, with majority metallic oxides. The TiO_2 $\text{P}43\bar{2}\text{nm}$ structure types (468 composition) are found directly below the CaF_2 compounds slightly to the right of the centre of the map in a small cluster, mostly containing intermetallic oxides. Going further towards the centre of the map gives the f-block intermetallic compounds with a p-block B ion, such as silicides and germanides, in the AlB_2 $\text{P}6\bar{3}\text{mmc}$ structure prototype.

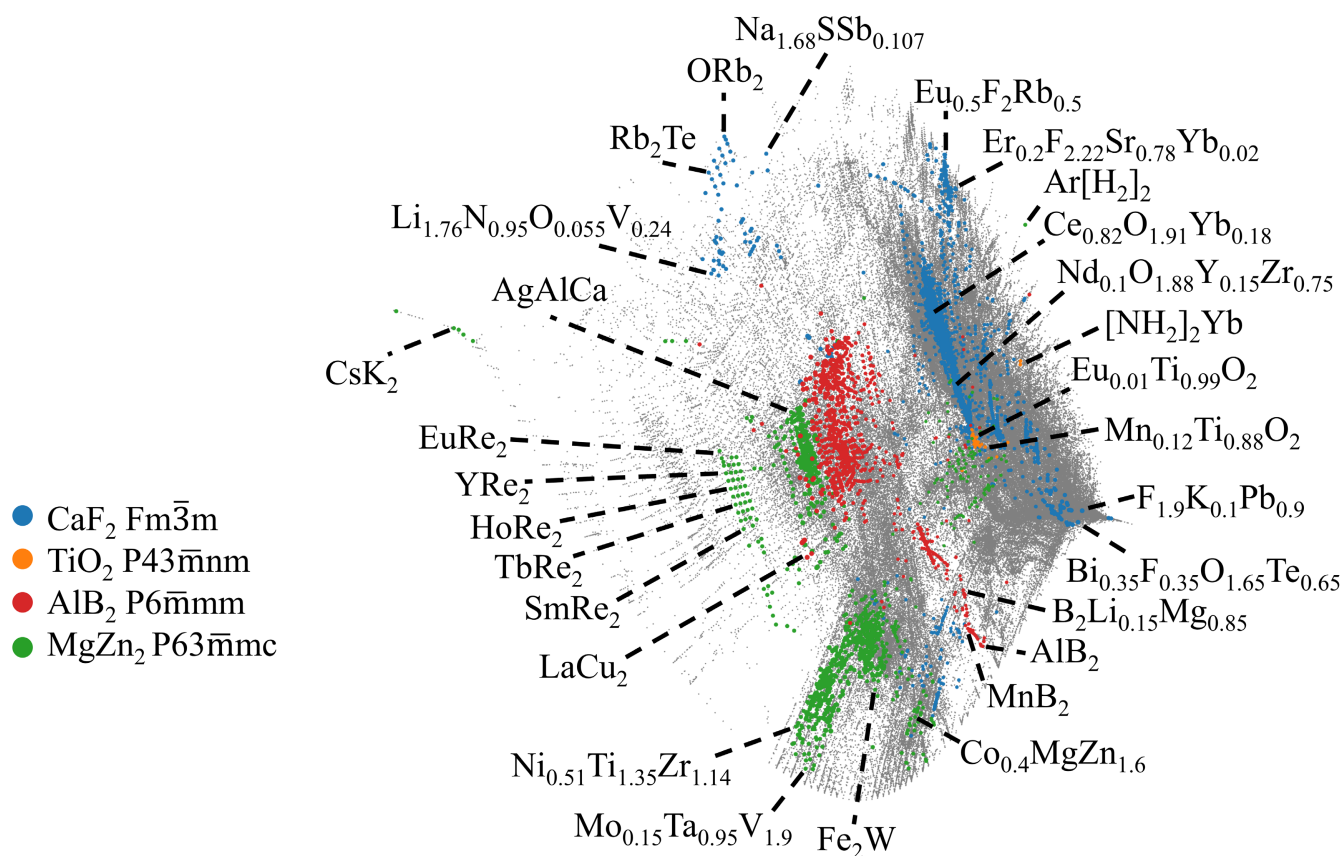


Figure 23: Different AB_2 structural prototypes may be found in distinct regions of the embedding in accordance with atomic weight of ions.

The composition AlB_2 itself is found to the lower right of the map at the intersection of two geodesics leading to Al and B. The f-block metal trends towards a d-block transition metal as we move from the centre, down the plot. Towards the lower centre of the map a selection of intermetallic transition metal compounds have the hexagonal $MgZn_2$ $P6_3\bar{m}mc$ structure prototype (1,078 total), with f-block and d-block intermetallic compounds with this structure type found in the centre of the plot.

As previously discussed, perovskite structures follow the ABC_3 ratio. In general these form cubic arrangements with where A site ions at the centre of the structure have a 12-fold coordination, B site ions found on the corners have 8-fold coordination, and edge centred C site ions bind these together. The undistorted perovskite structure has the prototypical structure of $CaTiO_3$ $Pm\bar{3}m$ (3,884 compositions). Separating these structures from the dataset in Figure 24 shows some of the type of formulations which lead to this structure.

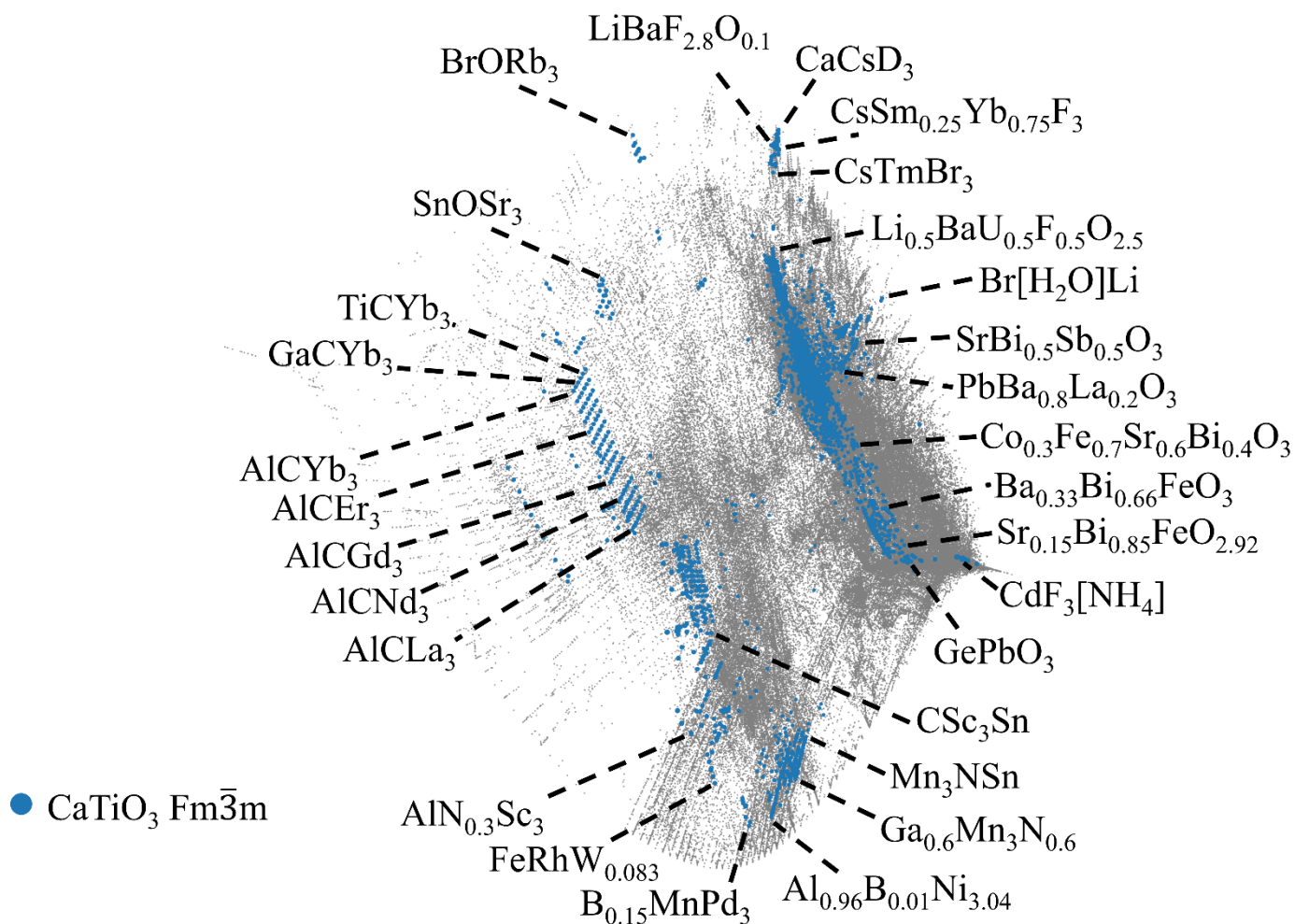


Figure 24: Each of the 3,884 materials in the Pearson's dataset with the CaTiO_3 structure prototype. These have clustered themselves into regions of ionic similarity. The ready availability of substitution in the oxide framework is reflected by the dense region in the upper right of the plot.

The lower region of Figure 24 shows the intermetallic transition metal perovskites. As the trend is followed up and to the left, we see that the inter-elemental electronegativity begins to diverge, with decreasing electron attraction on the C site until the anti-perovskite structures are found in a cluster at the upper left region of the map. These compounds have the same structure type, but a complete reversal of the charges at each structural site to the standard perovskite. Perovskites form stable arrangements of ions,

allowing for flexibility in the choice of dopants to create a new compound. The dense region of material in the upper right of the map almost entirely consists of oxides, showcasing a wide variety of dopants.

This feasibility of substitution has been well exploited by those who study the perovskite family, with many relying on an oxygen anionic framework. The $P4mm$ $BaTiO_3$, the $Pnma$ $GdFeO_3$, the $R\bar{3}c$ $LaAlO_3$, and the $R\bar{3}c$ $LiNbO_3$ structure prototypes are all distortions of the $CaTiO_3$ perovskite structure, which accommodates the differing ionic radii, Figure 25. Almost all of the 6,412 compounds assigned to one of these labels is located in the same region as the oxide $CaTiO_3$ structures, in the upper right portion of the map. Going to the bottom of this cluster of ABC_3 compounds gives intermetallic transition metal mixed anion perovskites with more covalent characteristics. The average A , and B ion electronegativity, decreases to the f-block metals, and the s-block metals as we move up the map. Following the right boundary to the top of the cluster gives many of the halide perovskites, with the f-block and s-block halides found in isolated clusters at the top of the map.

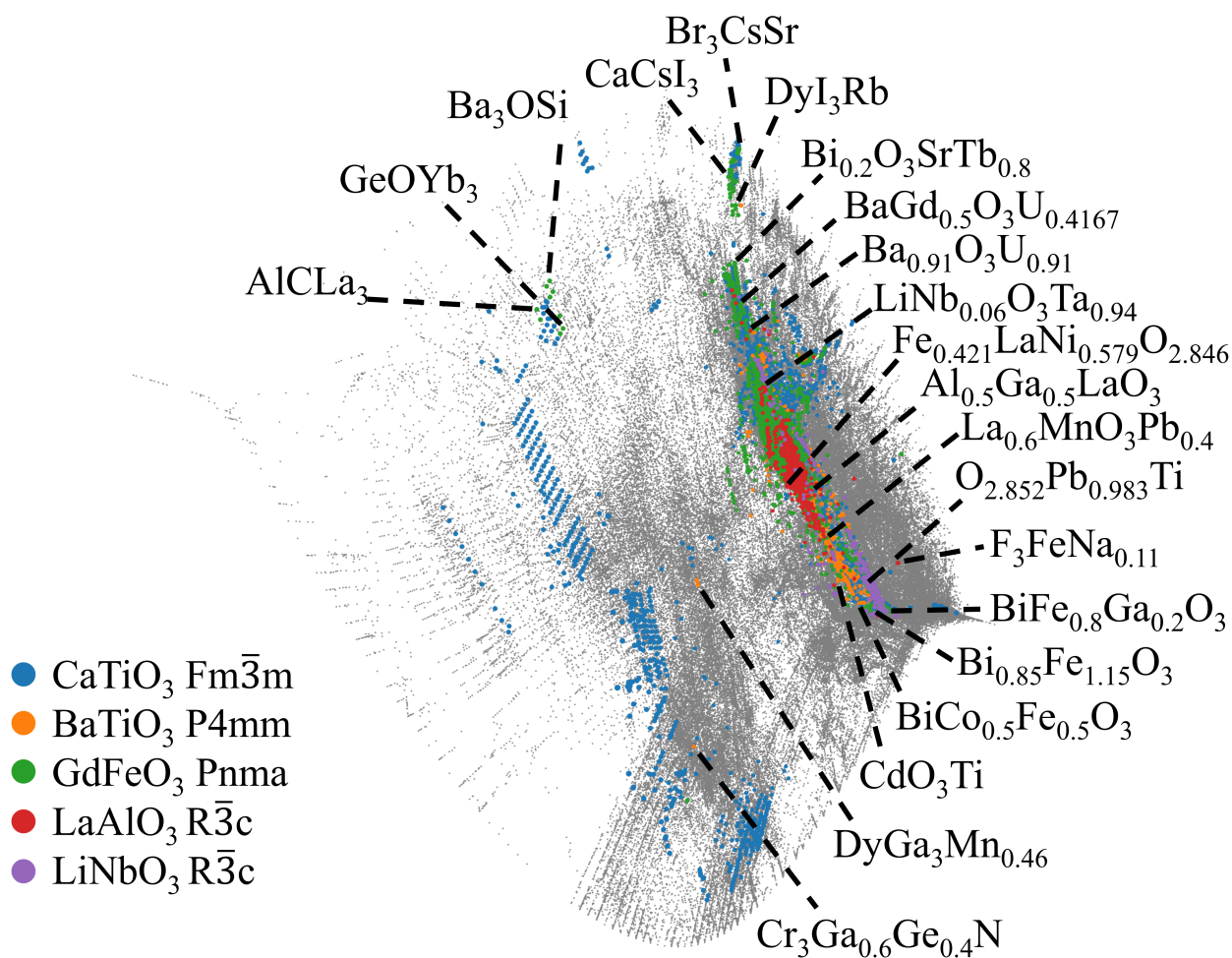


Figure 25: Overlaying common distortions of the CaTiO_3 perovskite structural prototype demonstrates some separation of materials based on chemical makeup, with the majority of compounds found in the same densely plotted oxide region of the embedding.

Conclusions

By directly calculating similarity of constituent elements, we present the ElMD as a computable mathematical relationship between any two compounds. This provides a natural extension to the physical scale introduced by Pettifor, allowing us to not only calculate the similarity of elements, but to quantitatively measure the similarity of compounds. These distances give a reliable measure of chemical similarity which align with human judgement,

which we may use to associate relationships and find patterns either analytically or with ML.

These distances present a method to connect separate chemical datasets and to pair compounds -- potentially theoretical ones -- to reported chemical information. A search interface using the compound formula as a query may be implemented, providing chemists with a natural interface to retrieve and explore data. This has been demonstrated by pairing a recent survey of 842 compositions with known ionic conductivity to their most likely reported structural information in the ICSD where, with a cutoff distance of 1, we have automatically returned a good match in 94% of cases. One clear future possibility is to connect chemical properties from multiple databases of potentially different compositions, where the distance may be used as a numeric measure of uncertainty for each assignment.

When designing statistical models, it is tempting to include all available chemical information in the hopes of arriving at the most accurate correlative results possible. There is however growing sentiment within the community to go further than simply black box curve fitting statistical models⁴⁵, with an increased call for interpretable models which not only give predictions, but also some understanding of how we have arrived at our answer. Here we use the EIMD to visualize and analyse solid state compounds in the ICSD, including the subset of binary compounds. With this metric we have created detailed chemical maps using modern data visualization techniques, which preserve clear trends in chemical relationships.

The quality of these maps is of high enough degree for the unsupervised ML method DBSCAN to automatically assign cluster labels such that similar compositions share a label. These assignments have a verifiable alignment with human judgement, which is given to us from the imbued domain knowledge engineered into the metric. Meaningfully understanding any large chemical dataset is a daunting task, and these maps aid us by

giving a broad overview of a compositional space. Employing the structural prototype labels provided by the Pearsons database exposes trends by reducing the space into subsets of materials with shared characteristics. This allows us to calibrate the metric against understanding, and confirms the relationship between chemical composition and resultant structural features.

It has been shown that simple metrics like the CED are ineffective for this, as they do not possess the resolution to differentiate disparate compositions in a space as complex as the domain of feasible compounds. This leads to an assessment of numeric similarity which does not align with chemical judgement, and in creating maps using this metric we find dissimilar compositions in close proximity to one another. In traditional ML models, for those with no background in statistical inference, determining why two points have a calculated proximity may be challenging. With the ElMD, should greater depth of investigation be required, a complete analytic solution can be calculated between two points to justify their exact positioning with respect to one other. These solutions provide chemists with thorough explanations for why two materials in a map have their calculated vicinity.

Understanding ML predictions requires us to not only understand the materials, but also the relationships between these. Stepping back from the forest of details allows us to look for general patterns, and with the results of ever more experiments readily available, we need ML to carry this forward. Following patterns to predict complex physical properties with 100% accuracy may prove to be impossible but we know that natural trends, although well hidden, almost always exist. If we are to understand these, we believe that the ElMD and other crafted metrics will prove to be invaluable tools in the categorization of materials space, and in further interpreting the AI we use in cheminformatics.

References

1. Schmidt, J. *et al.* Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem Mater* **29**, 5090 (2017).
2. Wen, C. *et al.* Machine learning assisted design of high entropy alloys with desired property. *Acta Mater* **170**, 109 (2019).
3. Jha, D. *et al.* ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **8**, 17593 (2018).
4. Rubner, Y., Tomasi, C. & Guibas, L. The Earth Mover's Distance as a Metric for Image Retrieval. *Int J Comput Vis* **40**, 99 (2000).
5. Orlova, D. Y. *et al.* Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE* **11**, e0151859 (2016).
6. Wu, L. *et al.* Word Mover's Embedding: From Word2Vec to Document Embedding. *ArXiv181101713 Cs Stat* (2018).
7. Monge, G. *Histoire de l'Académie Royale des Sciences de Paris.* (1781).
8. Pele, O. & Werman, M. A Linear Time Histogram Metric for Improved SIFT Matching. in *Computer Vision – ECCV 2008* (eds. Forsyth, D., Torr, P. & Zisserman, A.) vol. 5304 495–508 (Springer Berlin Heidelberg, 2008).
9. Király, Z. & Kovács, P. Efficient implementations of minimum-cost flow algorithms. *ArXiv12076381 Cs* (2012).
10. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. in *Advances in Neural Information Processing Systems* vol. 26 (Curran Associates, Inc., 2013).
11. Kolouri, S., Pope, P. E., Martin, C. E. & Rohde, G. K. Sliced-Wasserstein Autoencoder: An Embarrassingly Simple Generative Model. Preprint at <http://arxiv.org/abs/1804.01947> (2018).
12. Pettifor, D. G. A chemical scale for crystal-structure maps. *Solid State Commun.* **51**, 31–34 (1984).
13. Goldsmith, B. R., Boley, M., Vreeken, J., Scheffler, M. & Ghiringhelli, L. M. Uncovering structure-property relationships of materials by subgroup discovery. *New J Phys* **19**, 013031 (2017).
14. Isayev, O. *et al.* Materials cartography: Representing and mining materials space using structural and electronic fingerprints. *Chem Mater* **27**, 735 (2015).
15. Glawe, H., Sanna, A., Gross, E. K. U. & Marques, M. A. L. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New J. Phys.* **18**, 093011 (2016).

16. Ong, S. P. *et al.* Voltage, stability and diffusion barrier differences between sodium-ion and lithium-ion intercalation materials. *Energy Env. Sci* **4**, 3680 (2011).
17. Haghighatlari, M., Shih, C.-Y. & Hachmann, J. Thinking Globally, Acting Locally: On the Issue of Training Set Imbalance and the Case for Local Machine Learning Models in Chemistry. Preprint at <https://chemrxiv.org/engage/chemrxiv/article-details/60c745c4337d6cef32e2704f>. (2019).
18. Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D. & Rohde, G. K. Optimal Mass Transport: Signal processing and machine-learning applications. *IEEE Signal Process Mag* **34**, 43 (2017).
19. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and qsar modeling research. *J Chem Inf Model* **50**, 1189 (2010).
20. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).
21. Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W. Analysis and comparison of 2d fingerprints: Insights into database screening performance using eight fingerprint methods. *J Mol Graph.* **29**, 157 (2010).
22. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys Rev B Condens Matter Mater Phys* **87**, 184115 (2013).
23. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat Commun* **9**, 2775 (2018).
24. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Comput Mater Sci* **152**, 60 (2018).
25. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys Rev B Condens Matter Mater Phys* **96**, 024104 (2017).
26. Zhang, Y. *et al.* Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **10**, 5260 (2019).
27. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
28. Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *Npj Comput. Mater.* **8**, 1–26 (2022).
29. Hellenbrandt, M. The inorganic crystal structure database (ICSD) present and future. *Crystallogr Rev* **10**, 17 (2004).
30. Wood, E. A. Polymorphism in potassium niobate, sodium niobate, and other ABO₃ compounds. *Acta Crystallogr* **4**, 353 (1951).
31. Villars, P., Cenzual, K., Daams, J., Chen, Y. & Iwata, S. Data-driven atomic environment prediction for binaries using the Mendeleev number: Part 1. Composition AB. *J. Alloys Compd.* **367**, 167–175 (2004).
32. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension

- Reduction. Preprint at <https://arxiv.org/pdf/1802.03426.pdf>. (2018).
33. Zakutayev, A. *et al.* An open experimental database for exploring inorganic materials. *Sci Data* **5**, 180053 (2018).
 34. Hargreaves, C. J., Dyer, M. S., Gaultois, M. W., Kurlin, V. A. & Rosseinsky, M. J. The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions. *Chem. Mater.* **32**, 10610–10620 (2020).
 35. Villars, P. *et al.* Binary, ternary and quaternary compound former/nonformer prediction via Mendeleev number. *J. Alloys Compd.* **317–318**, 26–38 (2001).
 36. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
 37. Zhou, Q. *et al.* Learning atoms for materials discovery. *Proc. Natl. Acad. Sci.* **115**, E6411–E6417 (2018).
 38. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2**, 1–7 (2016).
 39. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).
 40. Schölkopf, B., Smola, A. & Müller, K. R. Kernel Principle Component Analysis. In *Proceedings of the Artificial Neural Networks—ICANN'97* 583–588 (1997)
 41. Dimensionality Reduction, Manifold Learning, and Metric Geometry. in *Topological Data Analysis for Genomics and Evolution: Topology in Biology* (eds. Blumberg, A. J. & Rabadan, R.) 235–270 (Cambridge University Press, 2019). doi:10.1017/9781316671665.006.
 42. Ohta, S.-I. Gradient flows on Wasserstein spaces over compact Alexandrov spaces. *Am J Math* **131**, 475 (2009).
 43. Villars, P., Cenzual, K., Gladyshevskii, R. & Iwata, S. Pauling File: Toward a Holistic View. in *Materials Informatics* 55–106 (John Wiley & Sons, Ltd, 2019). doi:10.1002/9783527802265.ch3.
 44. VLevy, A., & Lindenbaum, M. Sequential Karhunen-Loeve basis extraction and its application to images. *Proceedings 1998 International Conference on Image Processing. ICIP98I*, **2**, 456–460 (1998)
 45. Chuang, K. V. & Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chem Biol* **13**, 2819 (2018).

Supplementary Information

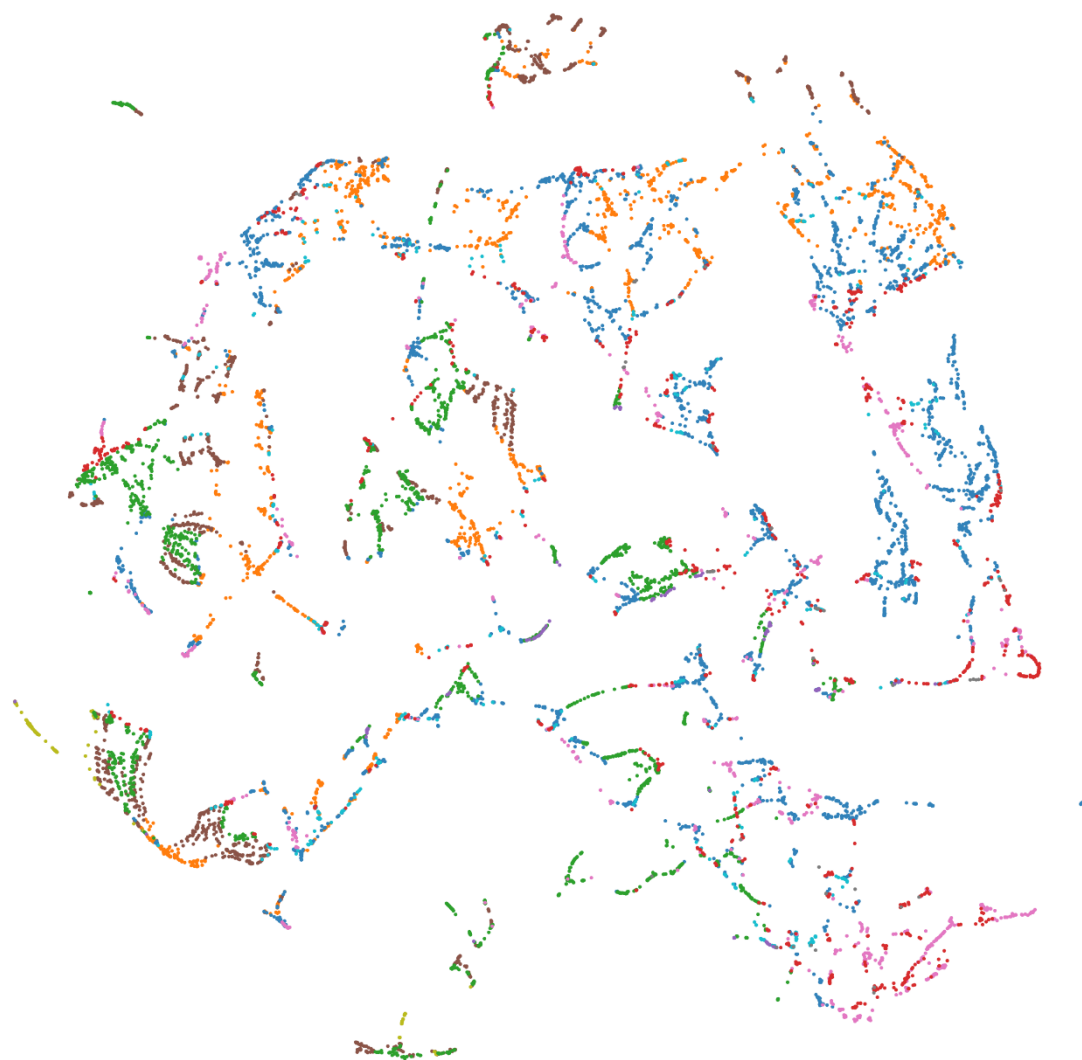


Figure S1: The atomic number is used to encode each element in an EIMD embedding of the binary compositions of the ICSD with UMAP.

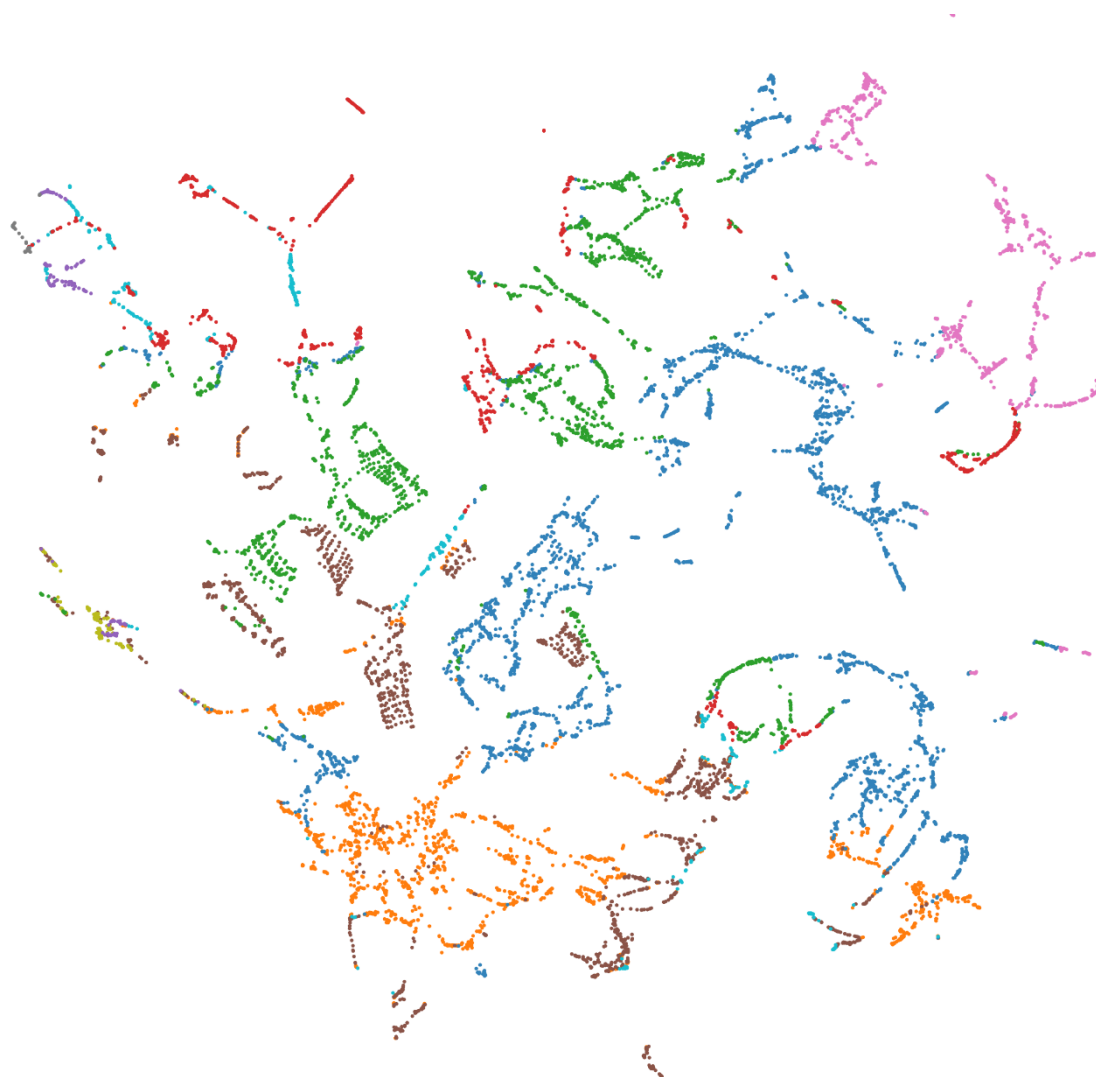


Figure S2: The Mendeleev number originally devised by Pettifor in 1986 is used to encode each element in an EIMD embedding of the binary compositions of the ICSD with UMAP.

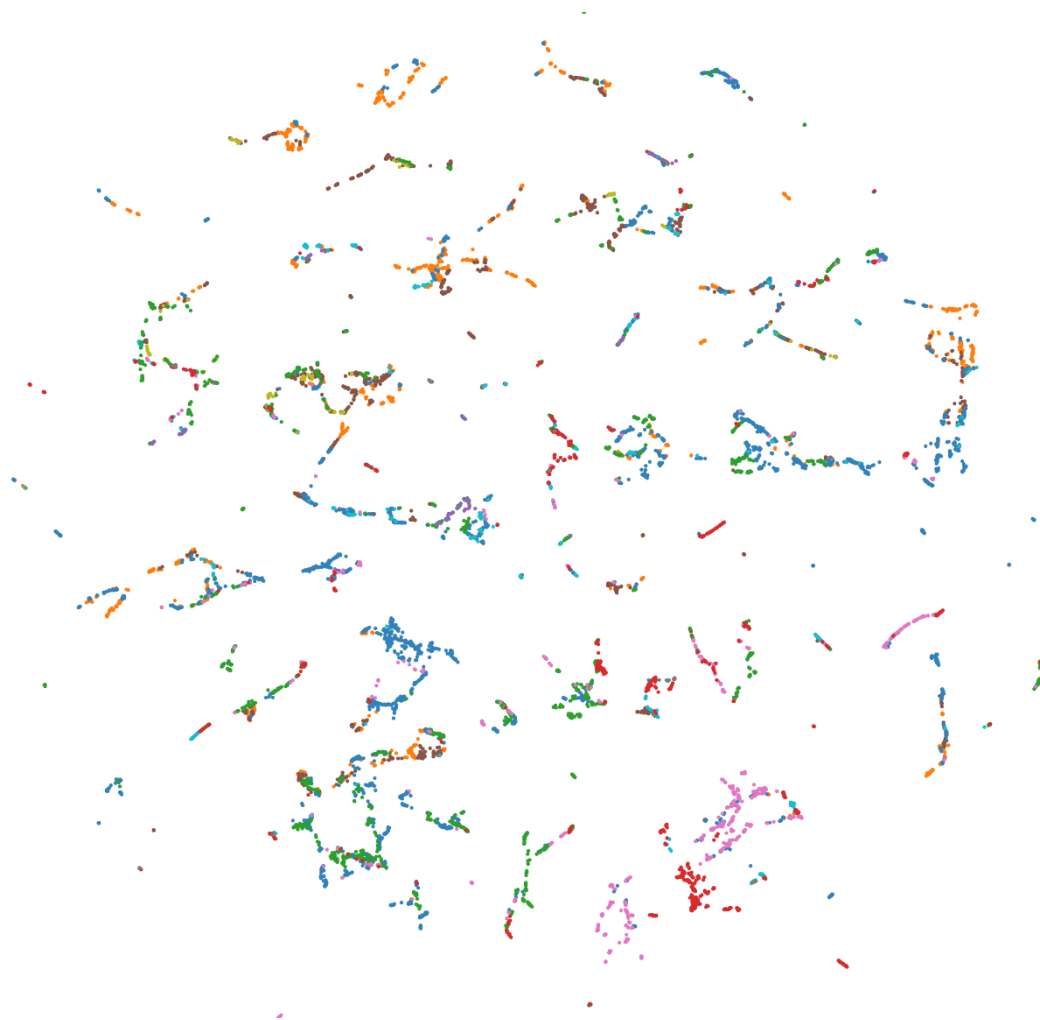


Figure S3: The Mendeleev number originally devised by Pettifor in 1986 is used to encode each element in an EIMD embedding of the binary compositions of the ICSD with UMAP.

Supplementary Note 1: DBSCAN analysis of binary compounds

With 12, 623 reported binary compounds reported in the ICSD the problem of organizing and categorizing these such that the families follow chemical reasoning is a tremendous task which is well suited to machine learning (ML). The application of the EMD to compositions gives us a reasoned measure of similarity between compounds, suitable for clustering.

Unfortunately, we may not use traditional clustering algorithms in the metric space directly, as the aforementioned negative curvature means that many of the Euclidean assumptions made by these algorithms, leads to nonsensical decision boundaries in the metric space. The application of the UMAP algorithm is therefore essential, and allows us to condense these relationships into 2-dimensional Euclidean coordinates which maintain appropriate local density whilst distancing from dissimilar neighbours.

Density based spatial clustering applications with noise (DBSCAN) is a popular clustering algorithm which is well suited to the embeddings produced by UMAP. At a high level, these labels are assigned by calculating the distance between all points, and any which are close by to one another are simply labelled as being within the same cluster. A distance threshold, Eps is assigned, and two points will be considered to not share a label if they are further apart than this. DBSCAN with an Eps-neighbourhood value of 0.63 on the UMAP embedded points initially creates 39 cluster labels, which when outliers of few compositions have been discarded, gives 26 core clusters.

These primary clusters are of small enough size that we may draw a reasonable qualitative classification, however several contain hundreds of compounds which may not immediately possess a trend which is discernible to the human chemist. Here follows an overview of each of the 26 clusters and the relevant chemical trends. In many of the larger clusters there is simply too much information for text to be an efficient method of communication,

and broader descriptions must be given. For ease of reference each cluster is projected to a 4×4 grid, and the corresponding grid references displayed in boldface. Global x and y intercepts have been overlaid, which due to the neighbourhood preservation given to us by UMAP, gives a weak equivalence to the EMD locally. The interested reader may find the full projection online at www.elmd.io/plots/binaries, where an interactive version can be further explored.

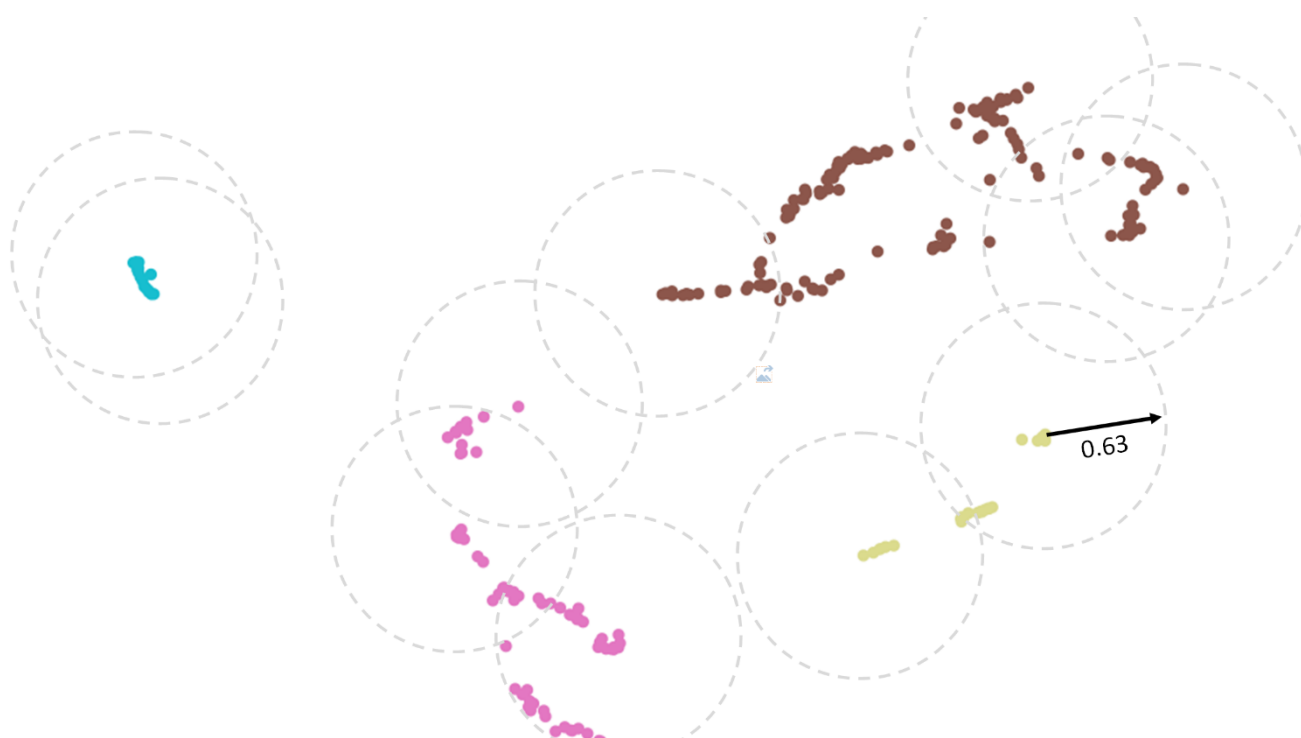
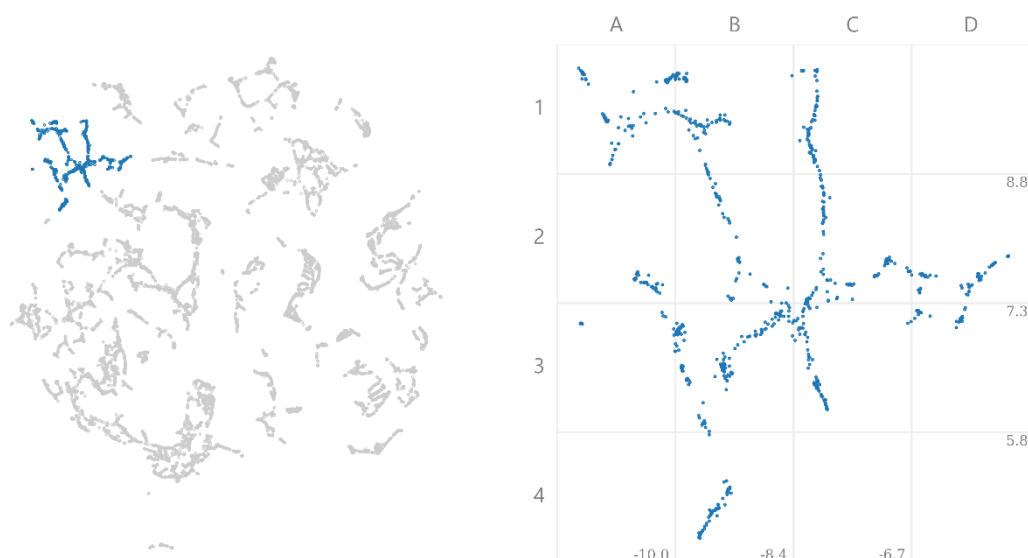


Figure S4: Clusters 14-17 of the binary compounds with disks overlaid on a selection of the cluster boundary points, to highlight the separation of these clusters from one another. In this example, there is an argument that due to their proximities and chemical similarities, clusters 14-16 could be considered to be the same cluster. If we increase the neighbourhood radius further to accommodate for this however, we see that major clusters elsewhere in the plot begin to merge also. For this embedding we have found that an Eps value of 0.63 has given good qualitative cluster labels overall, but the choice of this value will be entirely dependent on the dataset.

Table S1: The count and chemical labels present in each of the clusters assigned by DBSCAN with an $\epsilon=0.63$ on the 12,623 binary compositions in the ICSD (2017). DBSCAN was applied to the points obtained from UMAP on the binary compositions, reduced to two dimensions. The inter-compound EMD of each of these clusters was calculated, with mean, maximum, median, standard deviation, and local correlation of Euclidean distances with the corresponding EMD.

| Label | Count | Valence Bonds | Mean | Median | Max | Standard Deviation |
|-----------------|---------------|---------------|--------------|--------------|---------------|--------------------|
| Globally | 12,623 | | 23.09 | 21.68 | 101.16 | 11.56 |
| 0 | 843 | p-p | 6.18 | 6.17 | 16.57 | 2.81 |
| 1 | 335 | p-p, d-p | 5.44 | 5.14 | 15.17 | 2.76 |
| 2 | 262 | d-d, d-p | 6.07 | 5.83 | 19.25 | 2.98 |
| 3 | 295 | s-p, f-p | 5.94 | 5.74 | 17.03 | 2.96 |
| 4 | 516 | f-p | 8.59 | 8.58 | 24.06 | 3.89 |
| 5 | 40 | s-p | 4.84 | 4.44 | 11.96 | 3.24 |
| 6 | 914 | s-p, f-p | 10.28 | 9.83 | 33.93 | 4.87 |
| 7 | 223 | f-p | 4.84 | 4.58 | 14.71 | 2.74 |
| 8 | 830 | d-f, s-p, f-p | 11.21 | 10.50 | 49.75 | 5.96 |
| 9 | 290 | f-p, s-p | 9.05 | 8.83 | 23.86 | 4.61 |
| 10 | 250 | f-d, s-s | 10.39 | 8.60 | 41.59 | 7.56 |
| 11 | 688 | f-d, s-d, s-p | 12.11 | 11.58 | 35.15 | 5.81 |
| 12 | 204 | f-d, f-p, f-s | 7.56 | 6.25 | 26.50 | 5.29 |
| 13 | 496 | f-d | 7.23 | 7.00 | 22.00 | 3.58 |
| 14 | 222 | f-d, s-d | 4.42 | 4.32 | 13.71 | 2.11 |
| 15 | 32 | s-s | 2.10 | 2.21 | 5.57 | 1.36 |
| 16 | 27 | s-f | 0.65 | 3.05 | 1.86 | 0.46 |
| 17 | 178 | d-d | 3.11 | 0.57 | 8.48 | 1.64 |
| 18 | 40 | f-d | 1.41 | 1.28 | 4.90 | 0.96 |
| 19 | 425 | f-d | 5.58 | 5.55 | 14.67 | 2.41 |
| 20 | 2,729 | d-d, d-p, d-f | 10.20 | 9.71 | 48.30 | 5.25 |
| 21 | 90 | d-f | 2.43 | 2.43 | 6.40 | 1.43 |
| 22 | 64 | d-p | 3.01 | 2.92 | 7.33 | 1.60 |
| 23 | 931 | d-s | 6.09 | 5.40 | 22.40 | 3.49 |
| 24 | 1411 | d-p | 10.84 | 10.62 | 30.69 | 4.84 |
| 25 | 288 | d-p, f-p | 3.81 | 3.52 | 16.00 | 2.30 |
| Average | 485.5 | | 6.28 | 5.95 | 20.45 | 3.34 |



Label 0

Consisting entirely of p-block binary compounds. Towards the center of the plot at **B3** we see compounds with a nearly equal AB ratio with some variance. Each of the trend lines from the center lead towards a higher ion doping in a reasonably smooth fashion. It can be seen that each branch has a reasonably strong separation into different chemical families amongst the p-block elements. In the upper left following the trend from **A1** to **B2** we find two non-metals (non-halogens) in the ratio $AB_2 \rightarrow AB_9$. Non-metals bonded to p-block metals are found in **A2** to **B4**, from $AB_9 \rightarrow A_2B_3$, with the trend to the center leading to binary non-metal compositions. Transition metals bonded to a pnictogen element in high doping are found in **C3**. The region covered by **C1** to **D2** is majority B and C containing compounds, which increase in these elements concentration to the rightmost tip to almost pure B with V_3B_{497} .

Chemical Families:

- Upper Left: Two p-block non-metals, non-halogens
- Lower Left: p-block + p-block
- Lower Middle: Transition metal + pnictogen
- Upper Middle: Group IV + halogen
- Right: p-block metal + highly doped B/C

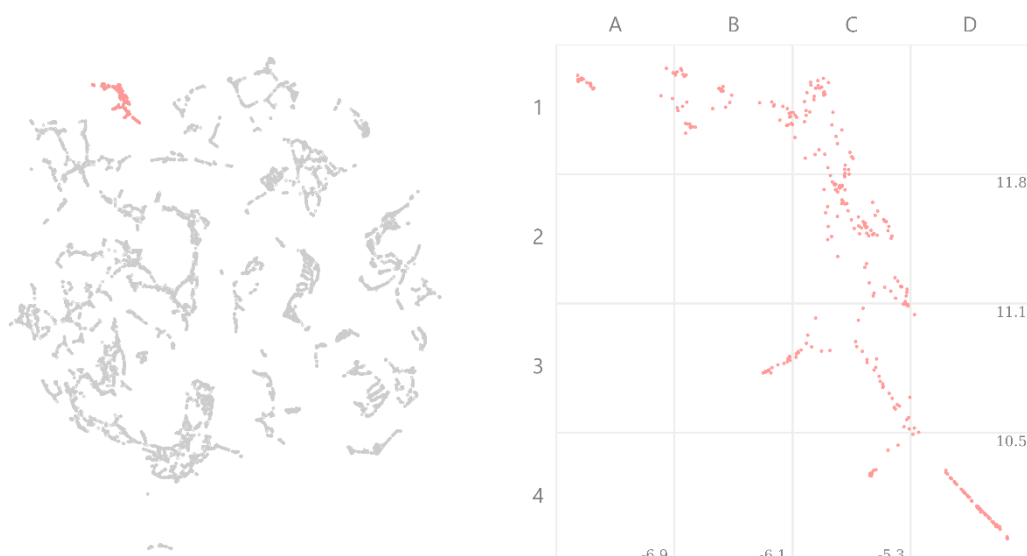
Composition Trends:

- Upper Left: $AB_2 \rightarrow AB_9$
- Lower Left: $AB_9 \rightarrow A_2B_3$
- Lower Middle: $AB_3 \rightarrow AB_4$
- Upper Middle: $AB \rightarrow A_4B_{121}$
- Right: $AB \rightarrow A_2B_{123}$

Chemical Trends:

- Upper Left: B ion increasing in atomic weight down chalcogen group. Partner ion follows metalloids plus carbon.

- Lower Left: Both ions increase in electronegativity from bottom to top.
- Lower Middle: p-block metal + non-metal, non-metal trends towards B/C as we move right
- Upper Middle: Carbon concentration dropping as we move up the cluster
- Right: Increase in B/C doping from left to right. Some silicides and germanides present in the center of the cluster.



Label 1

An island of Te/Se sulfides found to the left of **A1** leads to Te/Se iodides in **B1** before the remaining chalcogen halides from BiI to OH₂ are found in **C1**. Moving down the larger cluster we find the *A* ion moving through the metal-halides from PoF₅ to InF₃ in the general form *AB*₃. Reaching the split in **C3**, towards the left are B/C fluorides and hydrides, from BF₂ to B₉H₁₁. Following the branch to the right we see the p-block metal-halides from GeF₂ to PbI₂, until the lower split which follows boron group-halides to Ga₂I₃ in **C4**, and transition metal-halides of the form *AB*₂ from HgI₂ to PdI₂ in **D4**.

Chemical Families:

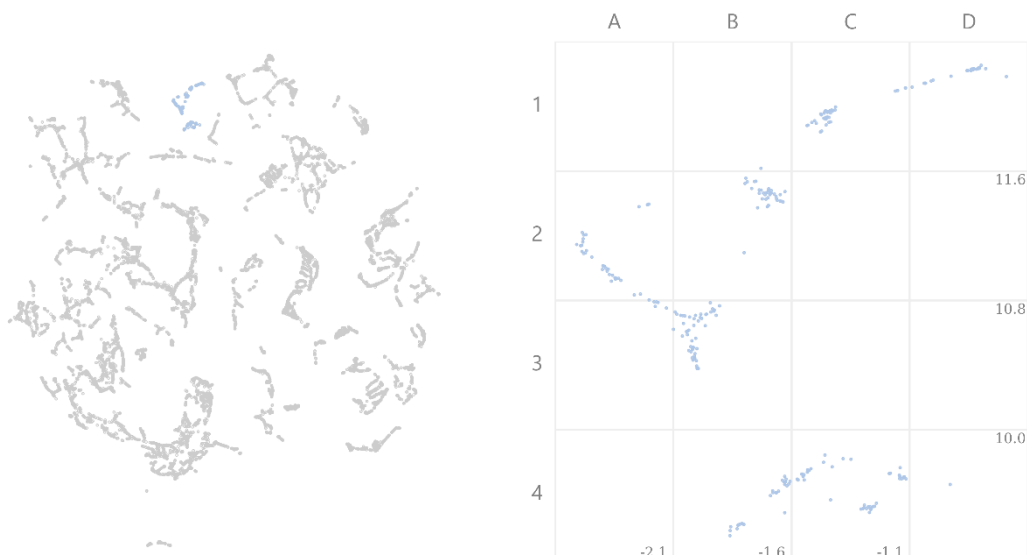
- Upper: Chalcogen + halogen
- Middle: Transition metal + halogen
- Lower: Transition metal + halogen

Chemical Trends:

- Upper: Large variance of *AB* ratio
- Middle: Generally of the form *AB*₃
- Lower: Trending to the left *AB*₂ → *AB*, to the right trends *A*₃*B*₁₇ → *AB*₂

Chemical Trends:

- Upper: From chalcogen + chalcogen pairs to chalcogen + halogen pairs at the top.
- Middle: A ion decreases in electronegativity towards the carbon group moving down
- Lower: To the left, boron/carbon halides. As we move down, A ion decreases in electronegativity through the p-block metals.



Label 2

Here we see a collection of metal-halides, with metal-chalcogenides in **B4-C4**, which all possess a strong B doping of 70-83% concentration. The initial trendline through **D1** sees f-block metals bonded to F and H of the form AB_4 . Moving down we see three successive clusters with a smoothly increasing A ratio from **C1** to **B3** from OsF_5 to Ta_3Br_7 , with fewer elemental trends through the periodic table found in this region to accommodate for this smooth change in ratio. The trend from **A2** to **C3** follows transition metal-halides of the form AB_3 with the metal decreasing in electronegativity as we move down. The final cluster of points in **B4-C4** follows transition metal chalcogenides from V_3O_7 to UTe_5 .

Chemical Families:

- Upper: f-block metal + halogen
- Middle: Transition metal + halogen
- Lower: Transition metal + chalcogen

Composition Trends:

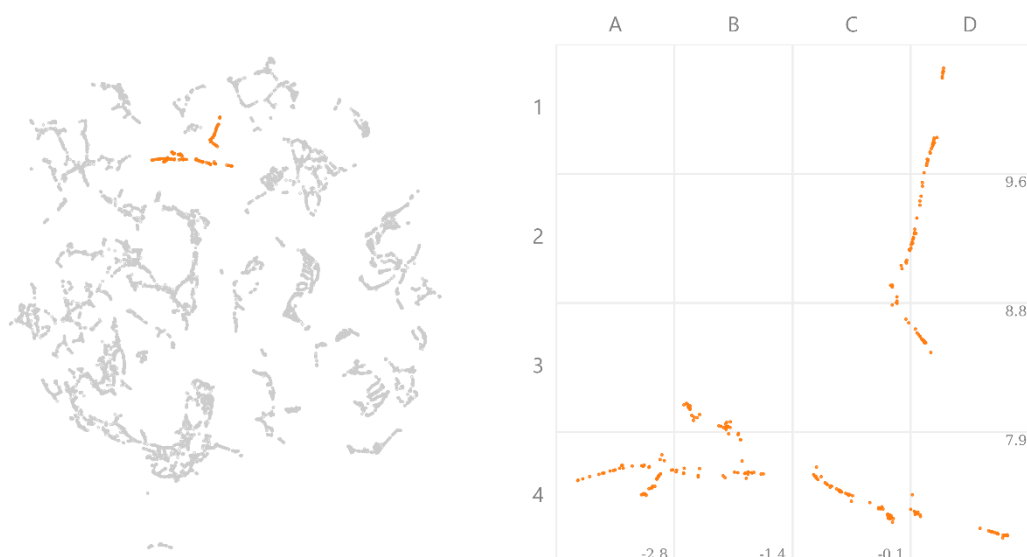
- Upper: $AB_5 \rightarrow AB_6$
- Middle: $AB_3 \rightarrow AB_4$
- Lower: $A_3B_7 \rightarrow AB_5$

Chemical Trends:

- Upper: From chalcogen + chalcogen pairs to chalcogen + halogen pairs at the top.

- Middle: *A* ion decreases in electronegativity towards the carbon group moving down

- Lower: To the left, boron/carbon halides. As we move down, *A* ion decreases in electronegativity through the p-block metals towards transition metal halides at the very bottom.



Label 3

The leftmost points of **A4**, with $\text{Na}_{0.021}\text{Si}_{0.979}$, follow the high *B* doping that can be found at the rightmost tip of cluster 0. This doping falls, but remains high as we follow the remaining alkali/alkaline metals along row **4**, mostly as silicides but with some other carbon group *B* ions, to SrSn_4 . The lower and upper branches along this trend, contain transition/f-block metal borides (**A4**) and carbides (**B3**) respectively. The line from **D1** to **D3** tends to follow f-block metal-carbon group compounds, from $\text{Ba}_{0.176}\text{P}_{0.824}$ to $\text{Nd}_{0.182}\text{Ge}_{0.818}$.

Chemical Families:

- Upper: Lanthanide + p-block
- Left: Alkali/alkaline/lanthanides + p-block
- Center: Lanthanide borides

Composition Trends:

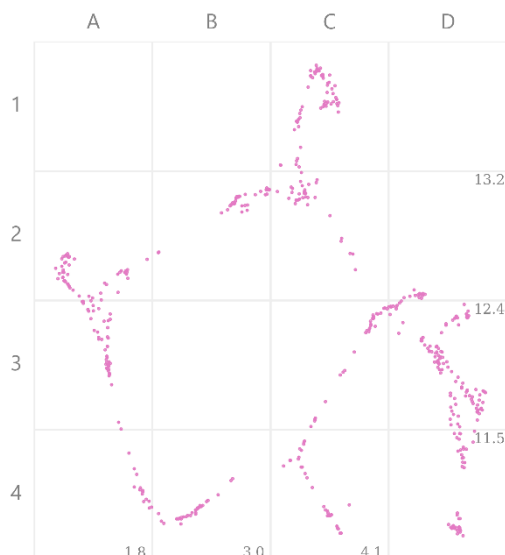
- Upper: $AB_7 \rightarrow A_3B_{17}$
- Left: $A_{0.021}B_{0.979} \rightarrow AB_4$
- Center: AB_4

Chemical Trends:

- Upper: Mostly lightly doped lanthanide carbides and borides, with greater variation at the top of the cluster, phosphides towards the center.

- Left: Mostly silicides, with a little variation amongst the metalloids combined with an alkali metal, with some alkaline metal/lanthanides.

- Center: Borides of the form AB_4 paired with an alkaline metal to the left, lanthanides to the middle, and light actinides in the far right cluster.



Label 4

This cluster possesses one of the clearest ring structures in the projection. Entirely consisting of ionic compounds (lanthanides and actinides bonded to a chalcogen/halogen B ion) this gives an interesting series of feasible substitutions and minor changes in stoichiometry, which will eventually return to the same compound. The upper and lower left regions of the ring are of the form AB_3 which smoothly trend to AB_2 compounds on the right of the cycle and the lanthanide-chalcogenides found in **D3-D4**.

Chemical Families:

- Left: Actinide/lanthanide/alkali + halogen/chalcogen
- Upper: Lanthanide/s-block + halogen
- Central and Right: Lanthanide/actinide + chalcogen

Composition Trends:

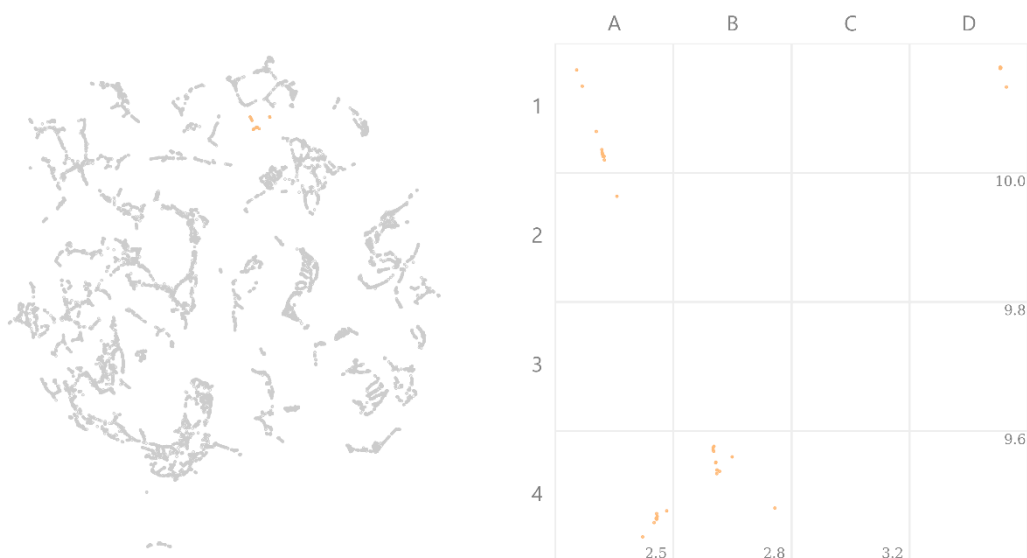
- Left: AB_3
- Upper Cluster: $AB_3 \rightarrow AB_2$, trace A_2B at top
- Central and Right: $AB_2 \rightarrow A_{0.433}B_{0.567}$

Chemical Trends:

- Left: Top left we have A ions as actinides in the form AB_3 which transition through the lanthanides, then alkaline metals combined with a halogen. After this, a B chalcogen, with A moving back through the alkali, lanthanide and actinides towards the center of the cluster.

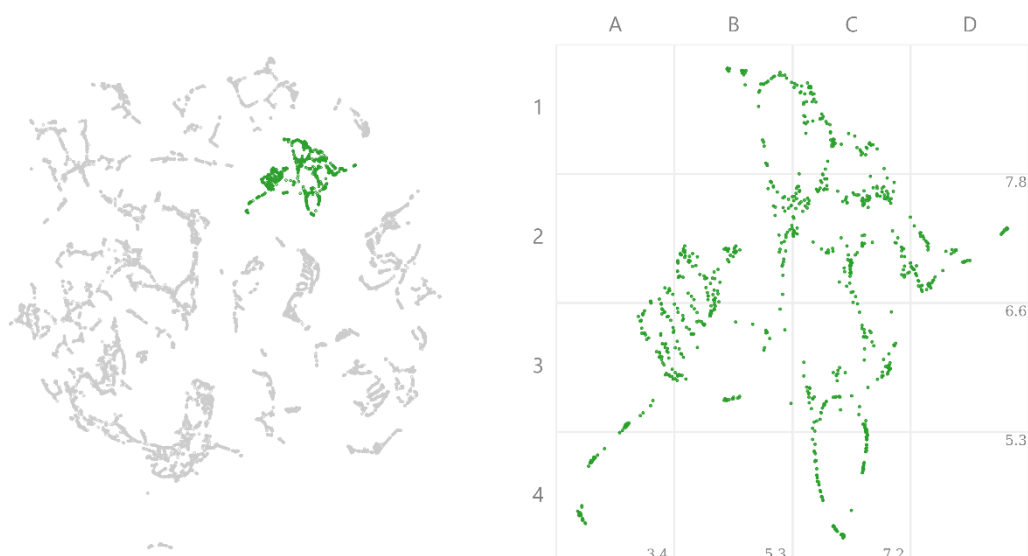
- Upper: Starting with a lanthanide and halogen on the left of this region, trending to an increase in B ratio as we go towards the center of the cluster. This splits off into the s-block metals towards the top of the cluster in the form A_2B

- Central and Right: Follows the trend of the left hand cluster in composition and makeup, splitting into two branches, for lanthanides and actinides, each paired with a chalcogen, of the form $A_{0.029}B_{0.71}$ and AB_2 respectively. As the upper cluster merges with the central cluster the A ratio continues to drop to the bottom right, with exclusively lanthanide A ion with a chalcogen



Label 5

As with the previous family these are entirely ionic compounds, being an s-block metal combined with a p-block element of the form $A_{0.286}B_{0.714}$ to $A_{0.214}B_{0.786}$, with B ions being Bi/Sb (**A1**), P/N (**A2-B4**), Se/Te (**D4**). A ring structure can be loosely interpreted when viewed in conjunction with the previous plot, however due to the fewer number of reported compositions there are larger jumps in makeup and stoichiometry when we attempt to follow a cycle.



Label 6

Here we see a larger collection of ionically bonded compounds which possesses both cycles of similar compositions, as well as approximate parallel lines found from **A2-B3**. Each of these lines represent compositions of the form AB_3 , with each distinct line containing the same B ion. Across the entire plot from lower leftmost point to the far right there is a consistent transition in AB ratio from $PuAl_4$ to Rb_2O_3 . From top to bottom there is a general decrease in B ion electronegativity from As to In, as A ions go through the s-block to the f-block metals.

Chemical Families:

- Left: Lanthanide A ion + B ion from the heavier boron group elements.
- Central: At the top contains alkali metals and alkaline metal A ions, with a pnictogen B ion. Towards the middle, more alkaline metal and lanthanide A ions with the B ion trending towards the C/N/P groups.
- Lower: Boron group + lanthanides/actinides. These trends follow to the right, increasing in A ion ratio.

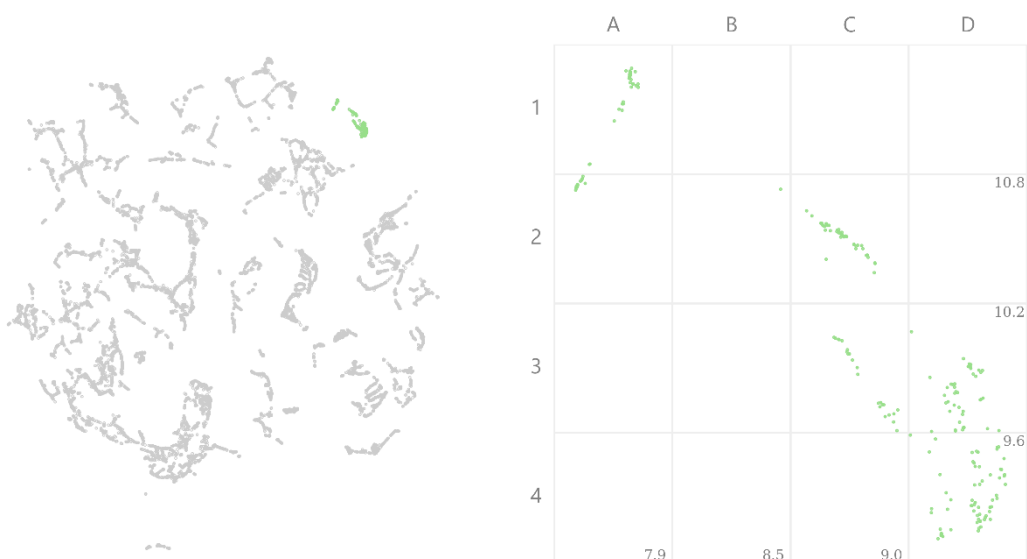
Compositional Trends:

- Left: $AB_4 \rightarrow AB_3$
- Central: AB_2 down the center of the cluster $\rightarrow A_2B_3$ to the right
- Lower: AB_2

Chemical Trends:

- Left: B ion ratio increasing from 0.2 to 0.32 to the center of the cluster. Characteristic parallel lines for AB_3 in the middle of the left hand, with each line consisting of a single boron group metal for each B ion. A ions trend from lanthanides in the far left, splitting into alkali/alkaline metal, lanthanide, and actinide in the center. These trends follow to the right.
- Central: As the trend moves further down the cluster, the B ion moves through the heavier carbon group, which tend to pair with an alkali or alkaline metal in the center. B ions tend to start in the boron group, trending to the right of the periodic table in the upper center of the cluster.

- Lower: B ion heavier carbon/boron group with heavier lanthanides



Label 7

The final cluster of ionic compounds for the most part continues the A_2B_3 ratio from the previous cluster, for f-block metal-chalcogenides, found in **D4**, with Eu_2O_3 found at the bottom most point. There is some separation into respective B ions through the chalcogen family as we follow the plot through each cluster up, with the inner crescent (**C3**) being of the separate ratio, $A_{0.429}B_{0.571}$.

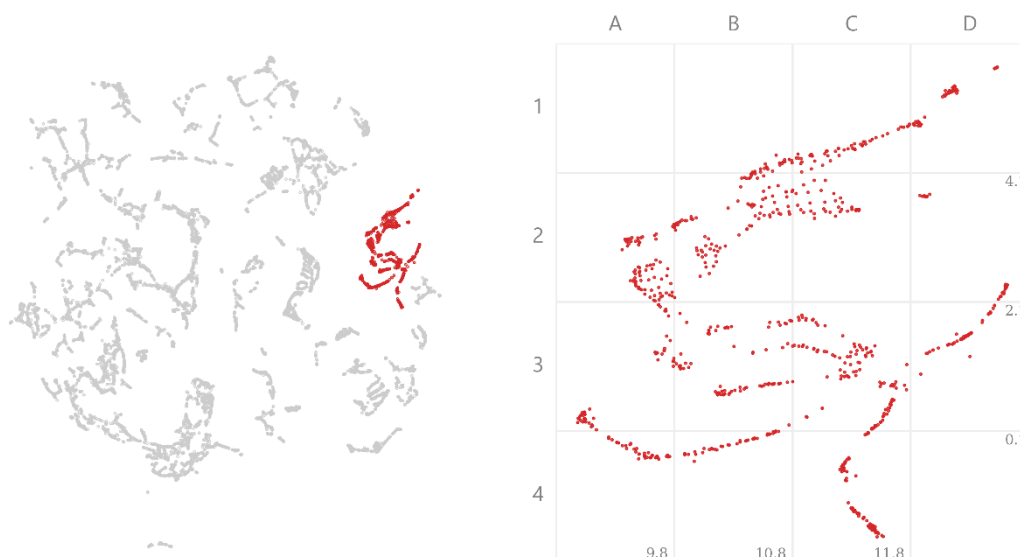
Chemical Families:

- Left: A collection of actinides and chalcogens with increased chalcogen ratio as we move down the cluster of the form A_2B_3 , where A is typically U/Th/Np/Am/Cm
- Right: From the upper left to the lower right the A ions follow the trend of the lanthanide series with a chalcogen ion, in small clusters separated by minor variations in ratio.

Composition Trends:

From left to right in two main clusters

- Left: $A_{0.429}B_{0.571} \rightarrow A_2B_3$
- Right: $A_{0.429}B_{0.571} \rightarrow A_2B_3$



Label 8

The majority of compositions in this cluster are AB compounds barring the lower branch (C4) which follows increasing A ion ratio to Nd_5Ge_3 . Here we see a reasonably wide variety of metallic elements from the s-block and f-block bonded with a similarly varied selection of p-block elements, reaching to the group 12 metals in the lower left branch. We see some regularity in the upper region where we may find a reasonably complete collection of AB lanthanide-chalcogenides.

Chemical Families:

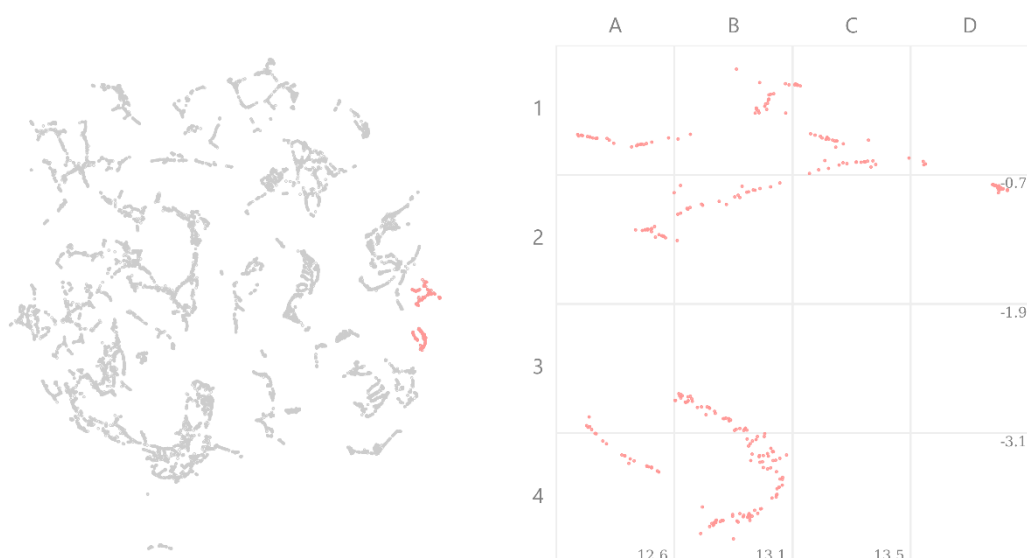
- Upper: Each of the lines moving up the cluster follow a B ion in the p-block. A ion trends towards the left of the transition block metals.
- Lower: metallic A ion with p-block B ion with

Composition Trends:

- Upper: AB
- Lower: AB , some trending $A_5B_4 \rightarrow A_5B_3$ in bottom branch

Chemical Trends:

- Upper: Increased chalcogen ratio as we move down the cluster of the form A_2B_3 , where A is typically U/Th/Np/Am/Cm
- Lower: Each of the arms from the center of this cluster show a classification in either composition or ratio. The center is AB compounds with high electronegativity, upper right contains alkali metals, lower right have divergent ratio from AB , bottom left are the remaining AB compounds. Moving up the cluster we see a split into multiple lines, each following a period in decreasing atomic weight as we move up.



Label 9

Here we follow the increasing A ion ratio from the previous cluster, reaching to Na_2S to the far right, with Sm_2O at the bottom of the plot. We see the typical trend from lanthanides to alkaline metals and alkali metals, repeated for each of the respective compositional ratios found in each cluster, with La_3Al_2 at the bottommost tip of **B4**, where two DBSCAN clusters have been merged for brevity.

Chemical Families:

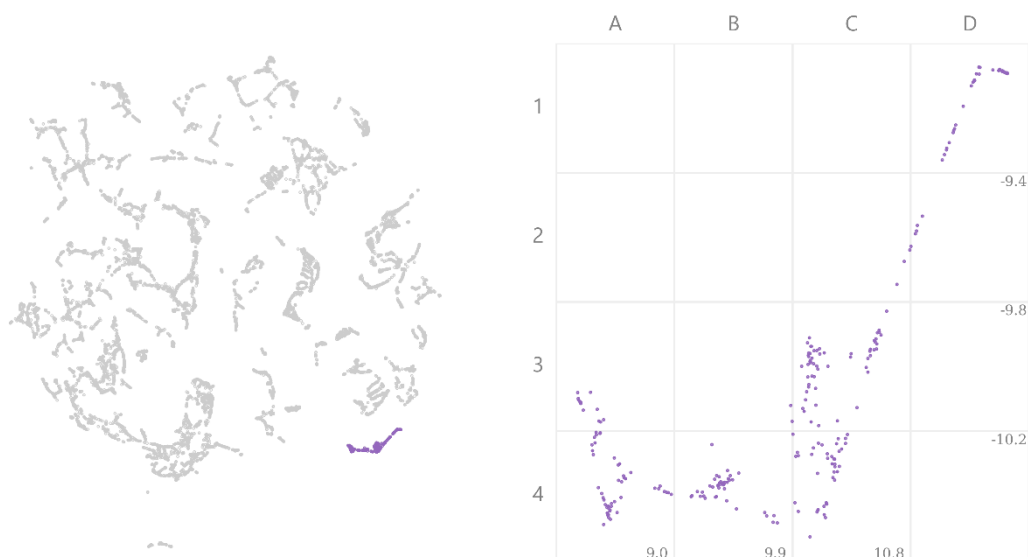
- Upper: Lanthanide/alkali/alkaline/actinides + boron/carbon group.
- Lower: Lanthanide/alkali/alkaline + Al/Ga \rightarrow p-block metals.

Composition Trends:

- Upper: $A_4B_3 \rightarrow A_5B_3 \rightarrow A_3B_2$ moving down.
- Lower: A_3B_2 and A_3B_5 in two distinct clusters.

Chemical Trends:

- Upper: No particular chemical trend barring those matching the clusters of ratio.
- Lower: From the upper left the B ion is Al/Ga, which transitions through the p-block metals in increasing electronegativity as we move down the cluster, trending through lanthanides to alkaline to alkali metal A ions moving down the cluster.



Label 10

The region in **C3** contains the reported binary compounds consisting of two f-block metals, with the trend to **D1** consisting of two s-block metals. The final cluster in **D1** consists of the XeCs compounds from XeCs₄ to Xe₃Cs. Following the trend to **B4** we see additional compounds containing Y, Sc, and Zr, with U/Pu-pnictogen binary compounds found in **A3-A4**.

Chemical Families:

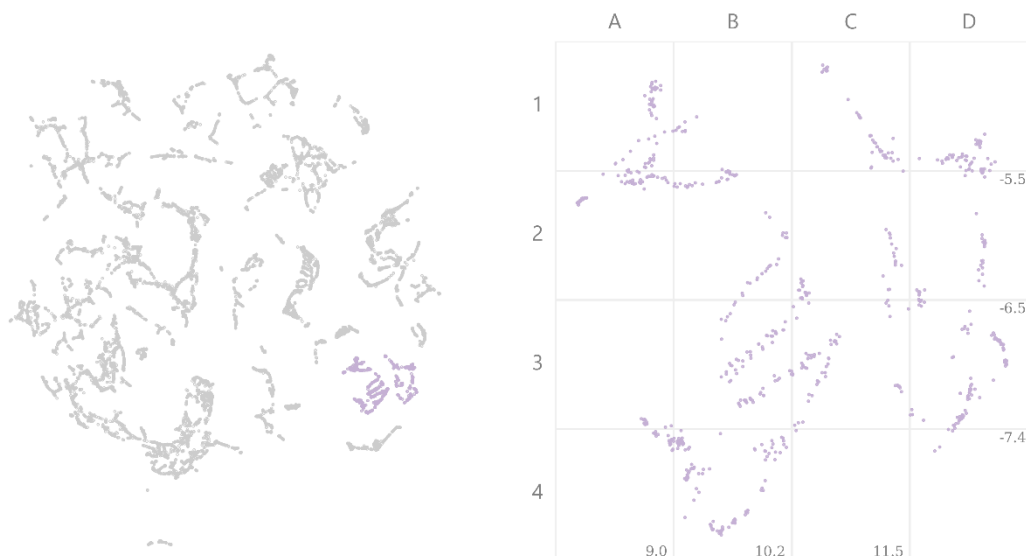
- Left: Actinide + p-block/transition metal
- Central: Lanthanide/actinide f-f compounds
- Upper right: Lanthanide + lanthanide in the center to alkaline + alkaline to alkali + alkali, transitioning smoothly between these.

Composition Trends:

- Left: Generally $A_{0.8-0.95}B$
- Central: Large variance, greater actinide concentration to the left of A_9B , and higher lanthanide concentration in the center of A_3B .
- Upper Right: Follows clusters of concentration through each of the chemical families, often to a high doping of one ion.

Chemical Trends:

- Left: Generally a high doping of U/Np/Pu/Th and a transition metal
- Central: Stronger ratios of actinides on the far left, generally transitioning through actinide for all the f-f binary compounds
- Upper Right: Moving up this cluster each of the main clumps center around high concentrations of elements from each group laid out above.



Label 11

Here we see a region of the map which due to chemical differences arguably should be separated into two clusters down the **C** column, but due to proximity have been classified under the same label with DBSCAN. The right hand side shows a natural continuation of cluster 9, following the same chemical families of s-block and f-block metals bonded to a p-block metal, with increasing *A* ion ratio moving down the plot from Th_2Ge (**C1**) to Ba_9Ga (**D4**). From right to left the *B* ion electronegativity decreases, which trend towards inter-metallic compounds, which are typical in the central and lower left regions of the global plot.

A1-B4 consists entirely of these inter-metallically bonded compounds, which are generally f-d valence bonded. From **A1** we find these in the A_3B_2 ratio with La_3Ir_2 , which increases in *A* ion ratio as we move around the cluster clockwise until La_3Sn at the bottom most point of **B4**, which come to more equal ratio to A_3B in **A3**. For each of the roughly parallel lines, these show variation in elements, with an f-block *A* ion and a d-block *B* ion, with each of the distinct lines possessing a fixed ratio, which are A_2B , A_7B_3 , and A_3B respectively. Each of the *B* ions throughout this cluster tend to be one of the following: Ir, Ru, Ni, Co, Au, Pt, Pd, Ni, Ti, Sn, Al, and Ge.

Chemical Families:

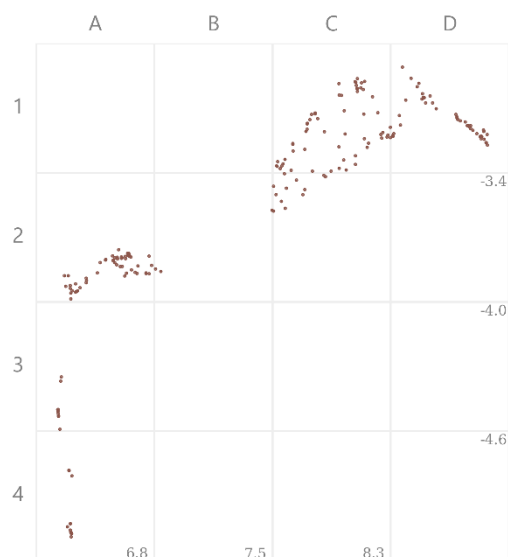
- Upper Left: Lanthanide + boron/carbon group *B* ion
- Left: Lanthanide + alkali/transition metal from the right of group 8.
- Right: Lanthanide/alkaline/alkali + p-block element

Composition Trends:

- Upper Left: A_4B_3
- Left: Each approximately parallel lines consist of a single ratio of elements, moving down the cluster these are $A_9B_{11} \rightarrow A_5B_4 \rightarrow A_3B_2 \rightarrow A_2B$ (first parallel line) $\rightarrow A_7B_3$ (second line) $\rightarrow A_3B$ (third line continuing to bottom of the cluster). The cluster to the bottom left trends from A_2B to A_3 .
- Right: $A_2B \rightarrow A_9B$ from top to bottom.

Chemical Trends:

- Upper Left: Lanthanide arsenides to the left, with B ion increasing in electronegativity as we move down the cluster.
- Left: Lanthanide A ion and a mixture of transition metals and metals ($Ru \rightarrow In + Mn$), B ion trends to the right hand side of the table as we move down the cluster. Moves through 9/10/11 B ion to s-block A ion from upper left to bottom right.
- Right: The entirety of the top of this cluster is of the form A_2B , as with the left hand cluster the A ion ratio increases as we move down this cluster, with the alkali/alkaline/lanthanide ratio increasing steadily to A_9B with $Ba_{0.909}Ga_{0.091}$.



Label 12

This region continues the chemical trends of the previous cluster, being majority f-d bonded intermetallic compounds of the same elemental makeup. The trend in B ion ratio is additionally preserved as we move further up, with almost all of these compounds being AB binary compounds, barring the trailing points in the lower left, which transition smoothly to A_3B as we rejoin cluster 11.

Chemical Families:

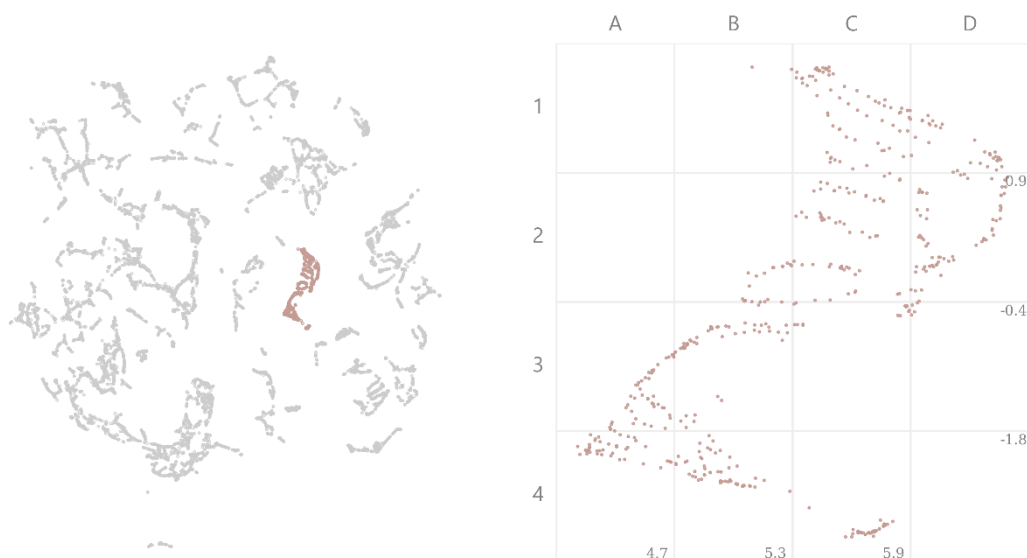
- Left: Lanthanide/actinide + boron group/transition.
- Right: Lanthanide/alkaline/alkali + group 10/11.

Composition Trends:

- Left: $A_3B \rightarrow AB$
- Right: AB

Chemical Trends:

- Left: There is heavier *A* doping towards the bottom of this cluster, with greater concentration of actinides. Moving up and to the right we see that the *B* ions become more metallic.
- Right: Lanthanide/alkaline/alkali + p-block element.



Label 13

Here we see the most prominent example of natural trends in substitutional feasibility being represented in graphical form. This cluster is almost exclusively AB_2 compounds, following a similar chemical trend to the previous two clusters. However with far greater numbers of reported stable compositions this allows for a more uniform arrangement of points which neatly follow trends in the Pettifor scale, see Figure 4 in the main body for a more thorough annotation.

Chemical Families:

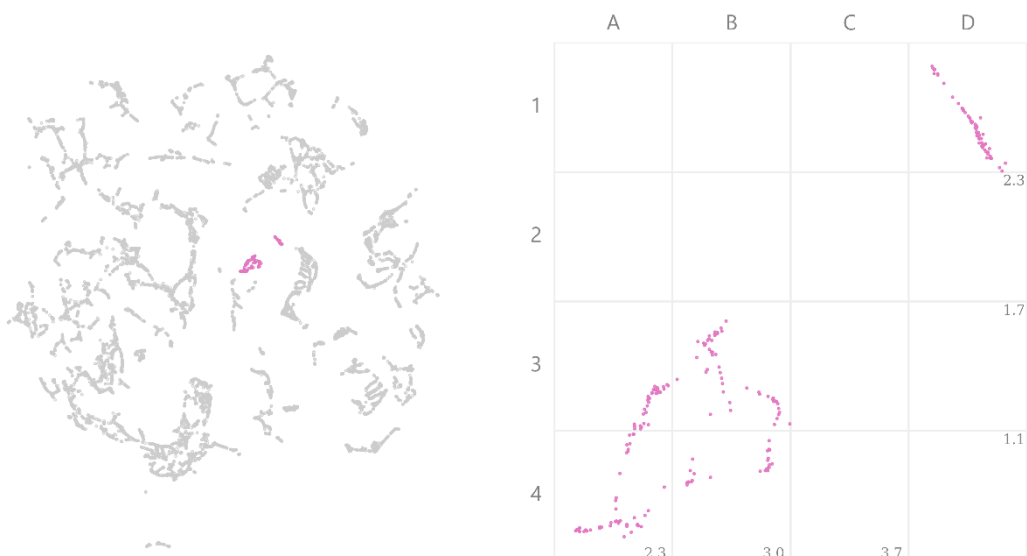
- Lanthanide/actinide + transition metal.

Composition Trends:

- AB_2 with some trending to A_3B_4 in the bottom right.

Chemical Trends:

- This follows a very strong ratio clustering, with almost all compounds being of the form AB_2 . The upper right of this cluster shows parallel lines, with each of these lines consisting of a lanthanides/actinide *A* ion, and a single transition metal *B* ion towards the upper right of the d-block. As we follow these lines to the east, these have the same *B* ion, with the *A* ion increasing in weight through the lanthanides and actinides. *B* ion trends from zinc to the middle of transition metals from top to bottom.



Label 14

The chemical trend of the previous group is again carried forward as a cluster of AB_3 ratio in **D1**. There are far fewer reported compositions with this ratio, and to the bottom left the trend continues to see a further increase in B ion stoichiometry, possessing the same elements found in cluster 13 but with less chemical ordering, from $Y_{0.207}Zn_{0.793}$ at the upper tip of **B3**, to $Nd_{0.143}Cd_{0.857}$ at the bottom left of **A4**.

Chemical Families:

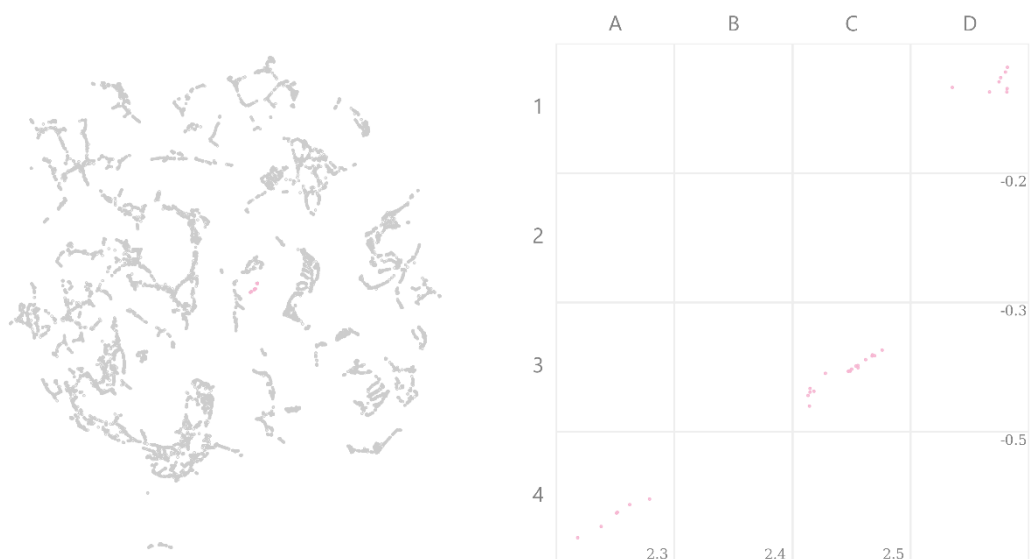
- Left: Lanthanide/alkaline/alkali + Zn/Cd/Hg
- Right: Lanthanide/actinide + Zn/Cd/Hg/Mg

Composition Trends:

- Left: $AB_6 \rightarrow AB_4$
- Right: AB_3

Chemical Trends:

- Left: A ion is generally lanthanide, with some alkaline, alkali and actinide. Trends to increased A ion ratio from left to right.
- Right: All of the form AB_3 following normal Pettifor trends with the noble gas inclusion of $XeMg_3$.



Label 15

This small cluster contains the entirety of the Mg metallicly bonded compounds with a *B* ion ratio greater than 79.3% going from $\text{Ba}_6\text{Mg}_{23}$ in **D1**, to $\text{Y}_{16}\text{Mg}_{109}$ in **A4**.

Chemical Families:

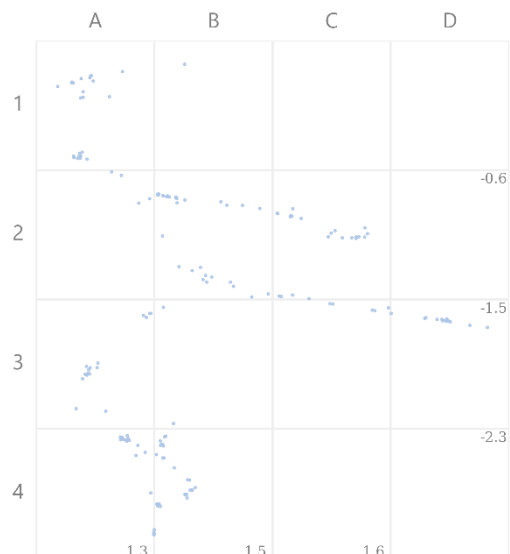
- Mg + s-block metal

Composition Trends:

- $A_{0.207}B_{0.793} \rightarrow A_{0.128}B_{0.872}$

Chemical Trends:

- Smooth variation of Mg doping from bottom to top of this cluster.



Label 16

We follow the continued decrease in B ion ratio moving further down the cluster, from AB_7 in **A1** reaching to $A_{0.077}B_{0.923}$ in **A3**, before following a small increase in A ratio with Fe/Mn bonded compounds to PrFe_7 in **A4**. The two branches follow Zn/Mg compounds with the ratios $A_{21}B_{179}$ and $A_{77}B_{923}$ respectively.

Chemical Families:

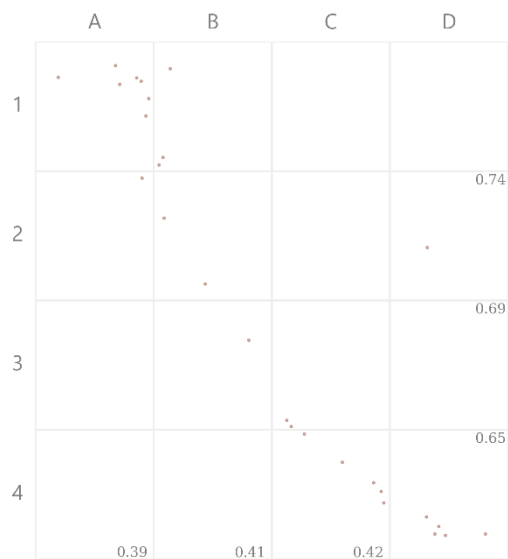
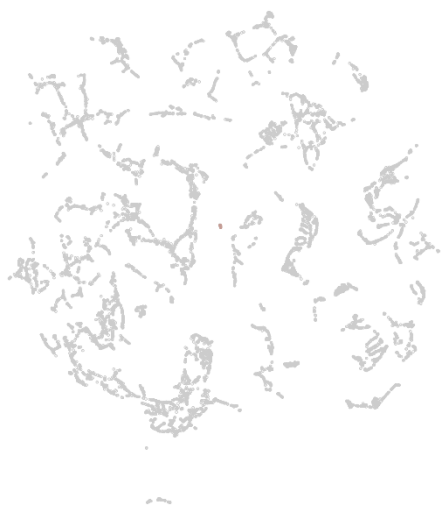
- f-block/s-block + Hg/Cd/Zn/Fe/Mn/Mg

Composition Trends:

- $AB_7 \rightarrow A_{0.07}B_{0.93}$

Chemical Trends:

- B ion goes to the left of the transition metals as we move down the cluster. A ion follows alkaline/lanthanide/actinide



Label 17

This small cluster contains the entirety of the Be compositions in the ratio $A_{0.071}B_{0.929}$ where the A ion is a lanthanide or actinide, which follows the same ratio as the compounds found at the closest point of cluster 16.

Chemical Families:

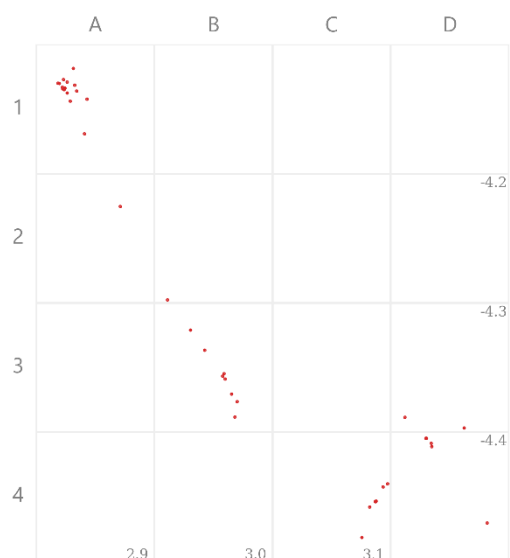
- Be + lanthanide/actinide

Composition Trends:

- $A_{0.071}B_{0.929}$

Chemical Trends:

- Cluster consisting of Be lightly doped with an f-block metal



Label 18

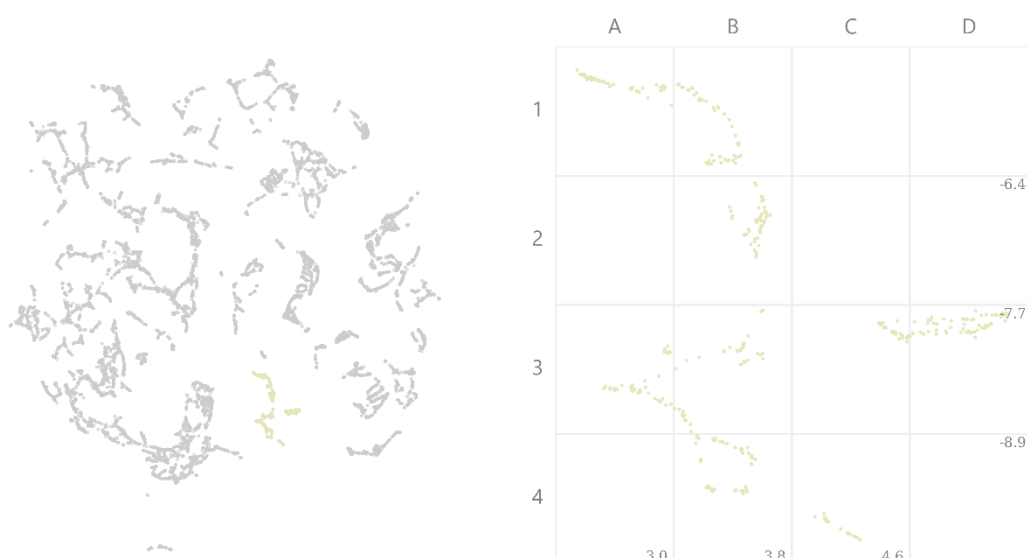
Here we have Mn (**A1-A3**) and Fe (**C4-D4**) binary compounds in the ratio $A_{0.207}B_{0.793}$ such that A is a metal from the lanthanide family.

Chemical Families:

- Lanthanide + Fe/Mn

Composition Trends:

- $A_{0.207}B_{0.793}$



Label 19

Here we follow the trend of ratio for the lanthanide/actinides bonded with a transition metal, however we see a discrepancy in the ordering. The top of this cluster consists of AB_3 compounds trending to AB_5 in **C4**, which if the previous trends had been observed would place this between clusters 13 and 14. There is little similarity with cluster 20 to the left to justify this placement, which is a known flaw of the UMAP algorithm when plotting “islands” of datapoints, as these clusters do not have similar enough compositions with neighboring clusters to place these entirely correctly. The overall chemical similarities with the previous clusters is still observed however.

Chemical Families:

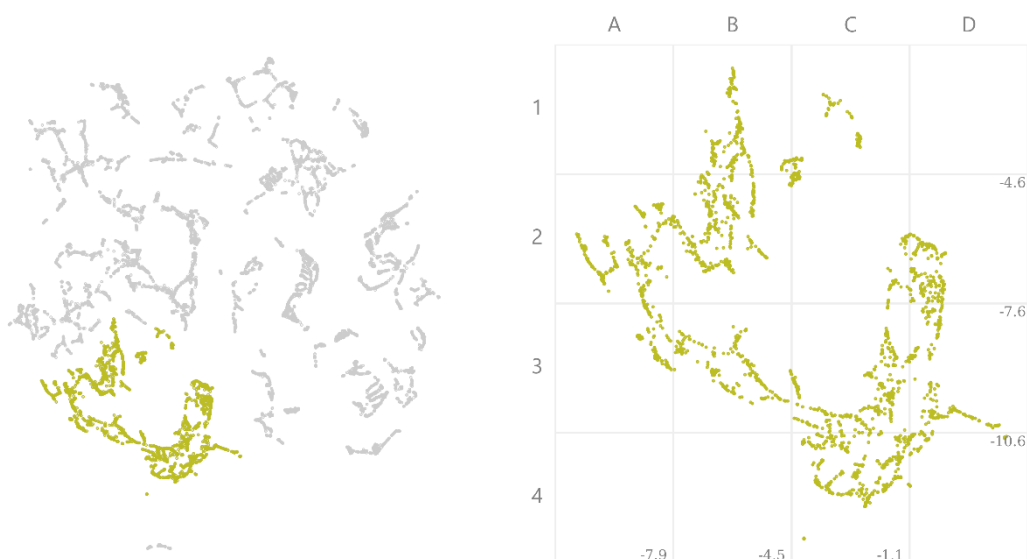
- Upper: Lanthanide/actinide + Ni/Co/Cu
- Right: Lanthanide/actinide + Rh/Ir/Pt/Pd/Au
- Lower: Lanthanide/actinide + Pd/Pt/Cu/Ag/Au

Composition Trends:

- Upper: $AB_3 \rightarrow AB_4$
- Right: AB_3
- Lower: $AB_{20} \rightarrow AB_5$

Chemical Trends:

- Upper: Going down the cluster, as B increases in proportion, it trends to the right of the transition metals, with actinides mostly found towards the center.
- Right: Transitions through group 9/10/11 metals, from left to right in B ion. A ion moves through actinides then lanthanides.
- Lower: Towards the center and left we see very high B ion concentration, towards the center of the d-block. A ion is a disordered collection of lanthanides, actinides and alkaline metals.



Label 20

The largest cluster, Of 2729 compositions, is invariably dominated by the transition metal compounds, which are almost exclusively found in this region. Here we see a complete range of ratios, and the trends as we move around the crescent tend to follow swells in ratio for specific elements, from AB compositions at **C3** to $DyCu_{99}$ at **A2**. There is a cluster of transition metals with more electronegative p-block elements at **D2-D4** which generally shows a higher transition metal doping, from A_2B to A_9B . The cluster from **B1-B2** leading up is initially two transition metals in A_3B ratio, before sharply increasing in transition metal doping as the B ion moves through to the p-block metals. This ratio evens out moving further up, leading to the next group of d-p bonded compounds.

Chemical Families:

- Right: Transition + transition
- Central Left: Transition + transition
- Upper Left: Transition + transition/p-block
- Upper Central: Transition + Zn/Cd/Hg/Be/Ga/Al

Composition Trends:

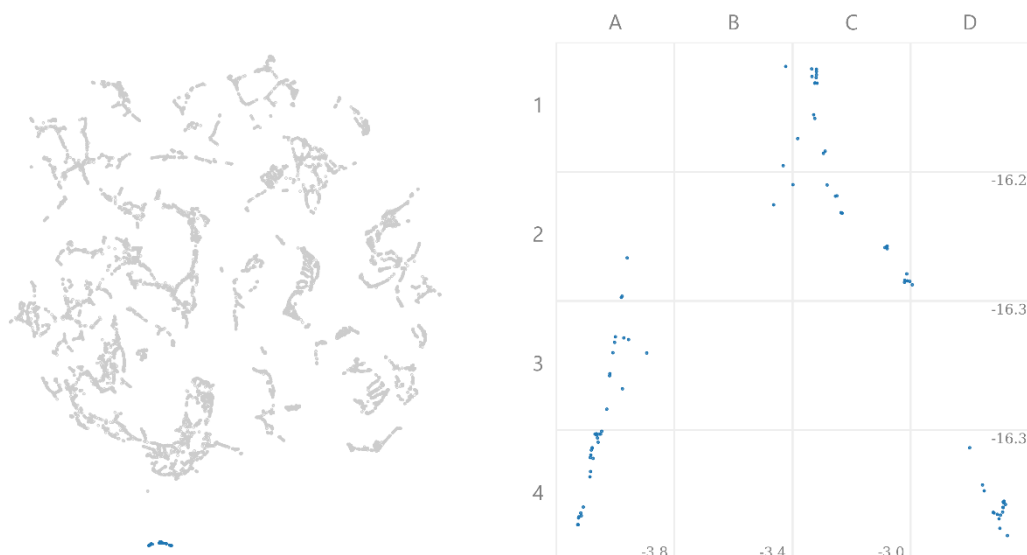
- Right: $AB \rightarrow A_3B$ from left to upper right. Lower cluster is mixture of AB , A_2B , A_2B_3 , A_3B , and $A_{20}B$
- Central Left: From left, high A ratio, $A_9B \rightarrow AB$ as we move down to the center.
- Upper Left: $AB \rightarrow A_4B \rightarrow A_2B$
- Upper Central: $AB \rightarrow AB_{99}$ and $A_3B_5 \rightarrow AB$

Chemical Trends:

- Right: Transition + transition
- Central Left: High ratios of Fe/Ni/Mn/Co at the left hand side of this region, transitioning smoothly in AB distribution to the center of the cluster. This more sharply trends to an equal distribution as it follows the main cluster up.

- Upper Left: Moving up the cluster, the B ion migrates to the right of the periodic table through the p-block metals. The B ion starts in small proportion, and increases as we move up the cluster.

- Upper Central: The left hand cluster are entirely Zn compounds increasing in concentration from left to right. The right hand cluster mostly contains binary compounds of transition metals from the very left of the d-block and p-block elements.



Label 21

These three isolated island each contain a distinct A ion (being Fe/Co/Ni) bonded to a lanthanide or actinide, in a reasonably wide range of ratios.

Chemical Families:

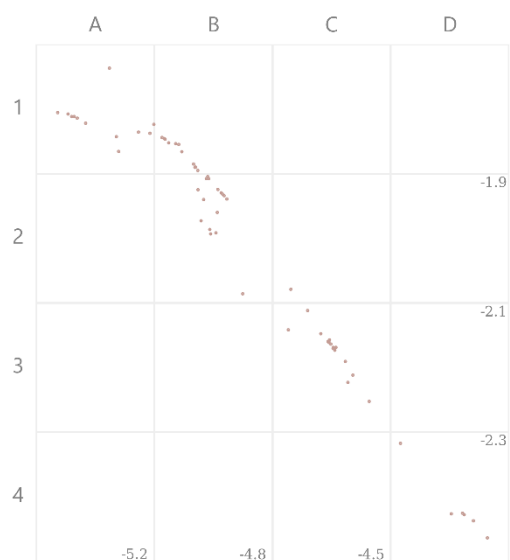
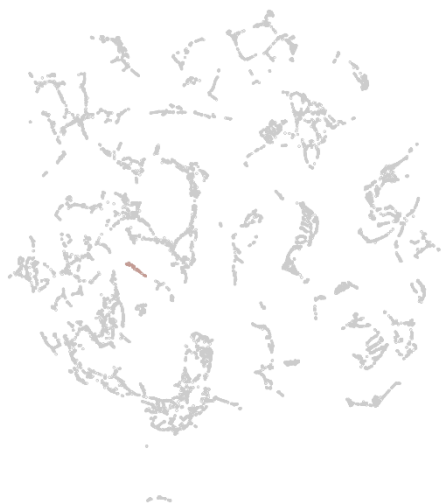
- Fe/Co/Ni + lanthanide/actinide

Composition Trends:

- $AB_5 \rightarrow A_{0.071}B_{0.929}$, disordered

Chemical Trends:

- Three islands each of a high concentration of Fe/Co/Ni, and separated by presence of this ion. A ion either a lanthanide or actinide with no discernable trend.



Label 22

Here we see a natural continuation of cluster 20 of d-p bonded compounds which have initially higher transition metal doping of AB_2 closer to the transition metal cluster, which leads to AB compounds as we go further towards **D4**.

Chemical Families:

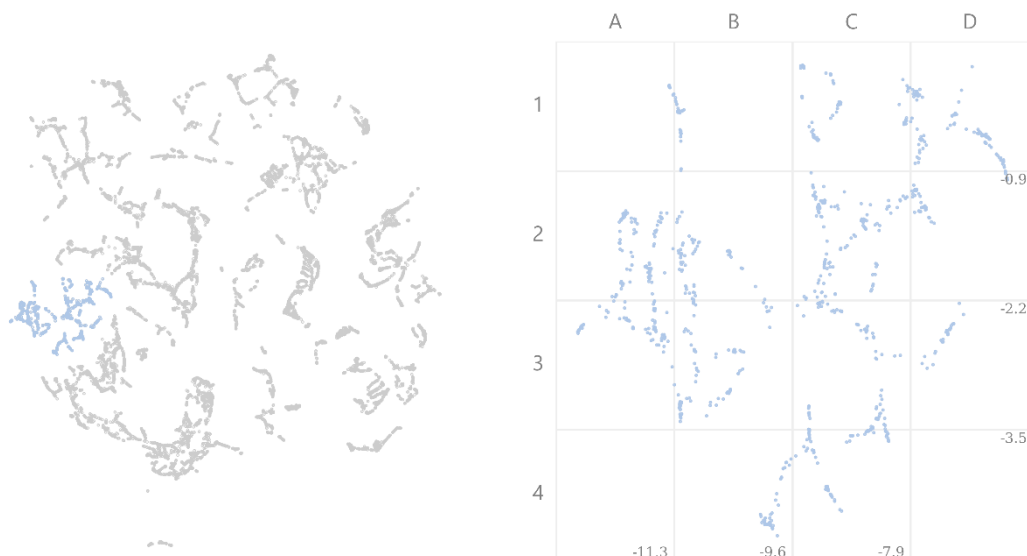
- Transition metal + Be/Al/Ge

Composition Trends:

- $AB_2 \rightarrow AB$, upper left to bottom right

Chemical Trends:

- As we move down the trend we see the left hand side transition metals plus Be/Al/Ga (in that order)



Label 23

We again continue the chemical trends of cluster 20 of high doping transition metal *A* ion with a p-block metal across **B4-D3**. As expected the ratio of *B* ions increases as we move towards cluster 0, with this being a post transition metal. Each of the branches tend to contain distinct *B* ions (Ga/Ba/Be/Al/In/Tl) and follow local trends in ratio, with AuTl₃ found in **D1**.

Chemical Families:

- Left: Transition metal + p-block
- Lower: Transition metal + Hg/Cd/Mg/Cu
- Upper Right: Transition metal + Mg/boron group

Composition Trends:

- Left: $AB \rightarrow A_{0.92}B_{0.08}$ increasing smoothly down lines, *AB* to the right
- Lower: $A_{19}B \rightarrow AB$ as we move from bottom to the center
- Upper Right: *AB* in the center. Trends to AB_{19} directly up, and to AB_3 diagonally to the right

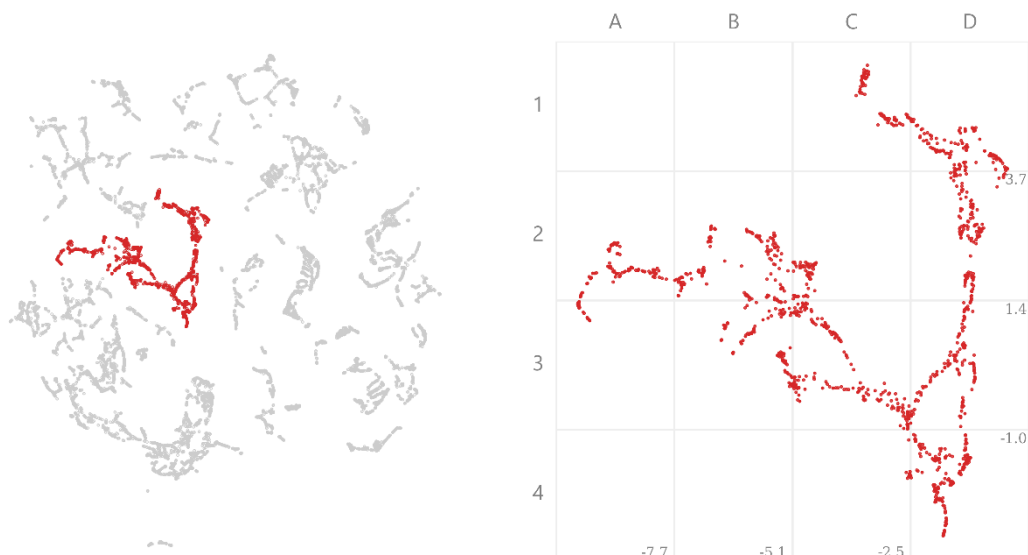
Chemical Trends:

- Left: High concentration of central d-block, mostly Fe and Co in the center, with branching into Ni/Co, Fe, Mn, Mg. The *A* ion trends follow the four roughly parallel lines down the cluster. The *B* ions tend to be from the boron, carbon and nitrogen groups, slightly increasing in atomic weight as we move to the right of the cluster. The top of the cluster is characterized by a halogen *B* ion with a transition metal in groups 9/10/11.

- Lower: Follows cluster to the left from *AB* increasing to a majority *A* ion ratio as we follow the cluster down. *A* is Cd, Mg, and Hg with a right hand side transition metal. At the very bottom this cluster the *B* ion moves further to the right of the table to give a selection of high Mg compounds.

- Upper Right: This follows the trend of increasing *B* ion ratio, with the *B* ion trending from magnesium to aluminum/beryllium. The transition metal trends further to the left of the periodic table as we move up the cluster. A

high concentration of metal towards the upper perimeters of the cluster, including Be/Al majority compounds.



Label 24

Here we find the second largest cluster of compositions, which is almost entirely transition metals bonded with p-block metalloids and non-metals. From **A2** to **D4** we find mostly AB compositions which trend to higher B concentration as we move towards **C2**. The B ion electronegativity increases as we follow the crescent up, through the pnictogens and chalcogens, with CoO_2 found in **C1**. Here we see that the branch at **D1** follows some of the halogens, suggesting that cluster 2 may have been misplaced by UMAP, as a more natural placement would follow to the right of this, before following on to cluster 3.

Chemical Families:

- Left: Transition metal/p-block + p-block
- Lower Right: Transition metal + p-block
- Upper Right: Transition metal + p-block

Composition Trends:

- Left: $AB_{19} \rightarrow A_3B_2 \rightarrow AB \rightarrow AB_2 \rightarrow A_2B_3 \rightarrow AB$
- Lower Right: $AB \rightarrow A_3B_2 \rightarrow A_5B_3$, going down
- Upper Right: $AB \rightarrow A_2B_3 \rightarrow AB_2$

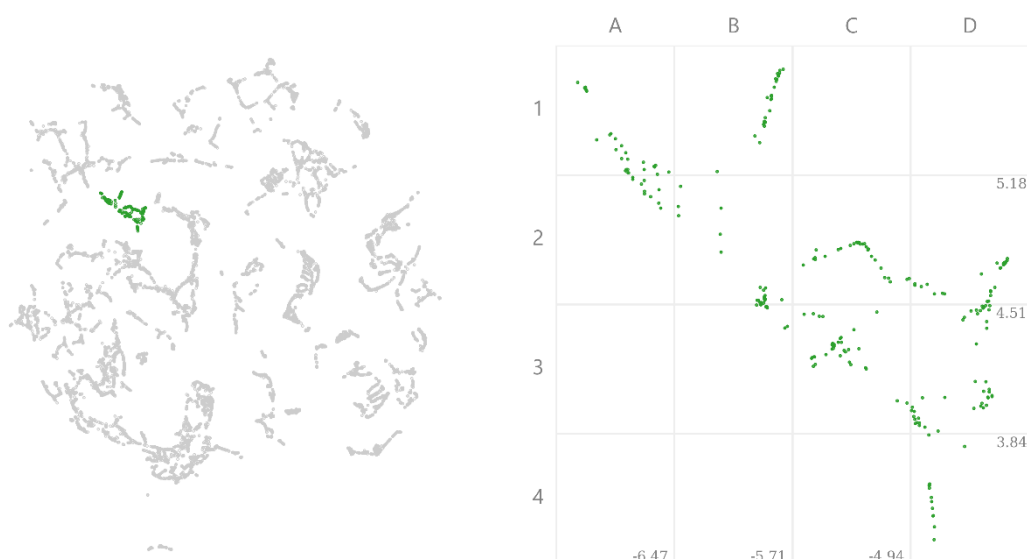
Chemical Trends:

- Left: This cluster begins with almost pure aluminum, with the dopants quickly becoming a d-block element in more equal ratios. This is initially a period 4 transition metal + p-block, but has group 9/10/11 metals + p-block as we move further to the right, or Mg/Be + p-block as we move further up. Towards the center are 9/10/11 metals + a lower left p-block. The A ions move to the left of the d-block as we keep moving to the right, and the B ion moves to the top of the p-block of the general form AB_2 . As we progress to

the lower right these return to AB nature with a left hand d-block and B/C/P/Si/As/Sb.

- Lower Right: As we keep moving down the cluster this increases in A metal content to A_5B_3 , of a left hand d-block + a carbon/pnictogen group. As we follow this up and to the right, the AB ratio remains more consistent, and the B ion moves towards the chalcogens (with a left hand d-block). Chalcogen content increases steadily as we move up the cluster to around $A_{11}B_{14}$ at the thinnest point as we move up

- Upper Right: Moving up, the B ion concentration increases to around AB_2 . This consists of mostly transition metals (plus some alkaline and lanthanide) A ion and chalcogen. Further up the cluster, the A ion trends to the right of the d-block. As we move to the far right of the cluster, we get the left hand d-block metals and a halogen equal ratio.



Label 25

The final cluster initially follows the trend from cluster 24, with transition metal AB chalcogenides found from **C3** to **D4**. The transition metal shows an increase in electronegativity as we move towards cluster 0 as we would expect. With AlO at **A1** we see the beginning of the p-p valence bonded binary compounds which makeup the entirety of the upper left region in cluster 0.

Chemical Families:

- Transition metal/actinide + chalcogen/halogen

Composition Trends:

- $AB \rightarrow A_2B_3$ in the far right

Chemical Trends:

- From the top left we have transition metals from the right of d-block A ions, in equal ratio to a halogen. As we move down and to the right, the A ion moves further of the left of the d-block to the lanthanides, with the B ion being a chalcogen element. Further to the bottom right we see a small increase in B ion ratio, which is mostly a chalcogen or pnictogen (often O)

A Database of Experimentally Measured Lithium Solid Electrolyte Conductivities Evaluated with Machine Learning

This chapter has been accepted for open access publication at *npj computational materials*, 9, 9 (2023), 10.1038/s41524-022-00951-z. I was the primary author with the much appreciated assistance of the co-authors Matthew S. Dyer, Michael W. Gaultois, Luke M. Daniels, and Matthew J. Rosseinsky. I lead the machine learning investigation under the supervision of the above, with the support of Vitaliy Kurlin. This involved continuing the previously unpublished work of Michael W. Gaultois and Emma J. Watts in gathering and validating suitable entries for the Liverpool ionics database. This required a co-ordination effort of many experimental chemists, which was driven by myself, Michael W. Gaultois, Luke M. Daniels, and Michael Moran.

We would like to thank the following listed co-authors on the submitted manuscript for their lengthy efforts in validating the entries of the ionics database: Beluvalli-Eshwarappa Prasad, Frédéric Blanc, Chris M. Collins, Catriona A. Crawford, Benjamin B. Duff, Jae Evans, Jacinthe Gamon, Guopeng Han, Bernhard T. Leube, Hongjun Niu, Arnaud J. Perez, Aris Robinson, Oliver Rogan, Paul M. Sharp, Elvis Shoko, Manel Sonni, William J. Thomas, Andrij Vasylenko, Lu Wang.

We would like to further express our sincere gratitude to the following listed co-authors, who undertook multiple rounds of data validation for the final stages of data cleaning: Michael Moran, Yun Dang, Rhun Morris, Alexandra Morscher, Kate Thompson, Matthew A. Wright.

Introduction

Energy storage is a key technology to meet growing energy demand by harnessing renewable sources. Liquid electrolyte-based Li-ion batteries have been extensively deployed in the portable electronic and electric vehicle markets. Alternative batteries that utilise solid state electrolytes (SSEs) avoid the safety issues associated with organic liquid electrolytes and offer high energy density by enabling the use of a lithium metal anode ^{1,2}. The most significant obstacle to the adoption of SSEs is the realisation of solid-state materials with the full suite of required properties, including sufficiently high ionic conductivity, stability against both Li metal and the oxidising cathode material (in practice this is often kinetic and associated with the formation of stable electronically insulating interfaces) together with appropriate mechanical properties ³. As such, considerable research has been devoted to the discovery and development of SSEs that meet these requirements ^{4,5}.

The amount of time and effort required to discover a suitable material in any domain has driven the application of machine learning methods to predict material properties ⁶. Recent works have used previously published data ^{7,8} to train machine learning models and predict the ionic conductivity performance of materials using only their composition ⁹. This approach is limited by the quality and quantity of the data available to train models. Literature reports in materials science tend to focus on subsets or particular families of materials with favourable or promising properties, leading to many reports on a limited range of materials ^{10,11}. While natural language processing (NLP) tasks have access to billions of training examples, in experimental materials science even large datasets typically contain fewer than 10,000 entries ¹². Due to these comparatively small training sets, it is imperative that the highest quality data are used to avoid providing inaccurate data to predictive models. As there are no large repositories of

experimental ionic conductivities currently available for solid Li conductors to perform a machine learning investigation, the first step must be sourcing high quality data.

Machine learning models for materials figure-of-merit performance can be built from knowledge of either the composition alone, or the structure and composition. While models built from knowledge of both structure and composition are generally superior in performance, composition-only models are important both for domains where we do not have structure and for specific considerations relevant to lithium ion conductors; the association between minor variations in stoichiometry and ionic conductivity, and the specific challenges in characterising these crystal structures.

The experimentally measured conductivity of a material derives from its non-averaged structure that is defined by its composition. This will include structural defects that cannot be captured fully in an average crystal structure recorded in a database such as the inorganic crystal structure database (ICSD), unless the material is fully ordered without fractional site occupancy or substitutional disorder. Most structures with lithium ion conductivity that have been reported in detail (*i.e.*, with the lithium positions) exhibit considerable disorder of this type. Even the average structure is unavailable for potential compositions that have not been experimentally studied, and in addition many experimental reports of ionic conductivity give composition but not structural analysis of the materials investigated.

Reported average crystallographic structures for lithium ion conductors frequently do not give precisely determined lithium positions because of the low X-ray scattering power and extensive structural disorder, again raising the important technical question of the connection between the potentially decisive local structure and the crystallographically-determined average structure. We thus build a dataset for machine learning models to predict lithium ion

conductivity based on composition. There will be limitations of this approach, for example the model will be unable to discriminate between polymorphs of a given compound. Nevertheless, crystal structure is not always known nor can it be for entirely novel compositions, thus a compositional model with low computational requirements is necessary for screening unexplored chemical space. Furthermore, many of these sensitively doped materials are highly dependent on fabrication methods, which can lead to broad inconsistencies across separate reports.

With few structures to hand, the experimental methodology taken in synthesis and property measurement reported by each source is the most direct method of judging each material. The most reliable measurement of the ionic conductivity of a material is via a.c. impedance spectroscopy (ACIS) measurement, usually on a dense ceramic¹³. This involves applying a voltage across the ceramic and sweeping through a range of frequencies. Measuring the voltage drop allows us to directly measure the complex impedance as well as the d.c. resistance of the material. All of the ionic conductivities for the materials included in this database were measured *via* ACIS.

For a specialist domain topic like solid electrolyte chemistry, the task of digesting the presented information requires significant expertise. Throughout the literature, there are inconsistencies in how data are presented, which introduces difficulties when comparing different reports. A broad knowledge of the background literature is essential for recognizing potentially problematic experimental procedures affecting both composition and conductivity, uncovering discrepancies in reported data, and identifying materials and properties that have in fact been computationally derived rather than experimentally measured (which problematically and unfortunately may not be clearly stated in the body of the text in some cases). All of these challenges increase the difficulty and time required to construct a high-quality database of experimentally reported data.

Leading natural language processing (NLP) approaches have demonstrated their capability to extract chemical data from the extensive corpus of past scientific literature ¹⁴, a process referred to as automated scraping. Text mining has been demonstrated to be a powerful tool in creating materials datasets. For example, Court and Cole ¹⁵ created a dataset of materials and their associated magnetic ordering temperatures. This is possible as a magnetic ordering temperature is reported as a single number usually in the text. Unfortunately for ionic conductors, the task of finding and pairing compositions, temperature of measurement, and conductivities is too complex even for state of the art NLP techniques to be effective. There are the standard issues of tokenizing chemical formulae consistently, and parsing correct values in text and tables. In particular for ionic conductors with a non-crystalline component, the composition is reported as a mixture of reactants rather than a stoichiometric chemical formula. Furthermore, as the vast majority of reported data is presented in figures with no standardized units for conductivity and extreme heterogeneity between entries, extracting relevant data is a combined challenge in both the fields of NLP and computer vision. Accordingly, the creation of a reliable database is unattainable with present automated capabilities, and thus a manual approach is employed here.

Previous investigations have predicted the ionic conductivity of solid-state materials using statistical methods. Due to the aforementioned difficulties in gathering initial datasets of sufficient size and quality these approaches build models that are based on relatively small experimentally-derived datasets (of the order of 40-82 entries).^{8,9,16} In this study, we have reviewed the literature to gather the largest currently available dataset of experimentally reported solid-state lithium ion conductivities. A statistical overview of the dataset is presented, with the range of conductivities examined for each structural prototype. Unsupervised embedding and clustering techniques are used to

partition this dataset into 9 families by compositional similarity, thus assessing the diversity of the dataset. We develop supervised regression and classification models to predict the lithium ion conductivity and assess whether a material will possess an ionic conductivity $\log_{10}(\sigma) > -4$ at room temperature, where the conductivity is reported in units of S cm^{-1} . The best regression models achieve a mean absolute Table 4 for $\log_{10}(\sigma)$ of 0.85, and the best classification models have a Matthews Correlation Coefficient (MCC) of 0.63, assessed under k -folds cross-validation in both cases.

Results and Discussion

Database Construction

A large collection of solid-state lithium electrolyte literature was gathered, and the ionic conductivities were extracted for the materials reported in each study. The experimental procedures in a given source were critically assessed to understand how each sample was synthesized, characterized, and processed into a ceramic. We ensure that in each of the studies, samples had clearly defined compositions and reported direct measurements of the conductivity taken via ACIS. The values of ionic conductivity in the database are a mixture of bulk and total values, as the two are not always distinguished, with only a small number of studies providing sufficient detail in labelling the reported values as such. Where exact stoichiometry may be unclear from the given reagents, any studies that lacked supporting characterization (such as ICP analysis) to confirm the presence of Li, were discarded. The ionic conductivity and material composition are both of equal importance in the database, as the predictive models are constructed with these two variables. By ensuring that data is exclusively gathered from experimental studies of high calibre, we gain confidence in the quality of the results of subsequent machine learning analysis. Typically, this requires extracting the values

from an Arrhenius plot and converting each value from the plotted units (commonly plotted as either σ in S cm^{-1} or S m^{-1} , $\log_{10}(\sigma)$, $\log_{10}(\sigma T)$, or $\ln(\sigma T)$) to conductivity in S cm^{-1} at a specific temperature. In some reports these values may also be provided in tables, or stated in the main body of text along with supporting discussion, allowing for cross-checking of the reported value.

The first stage of the initial literature review was carried out by an undergraduate student to collate source papers of reported conductivities from keyword searches using search engines, and reviews of the field ^{17–20}. This survey focussed on tabulating the physical properties reported in each paper: composition, ionic conductivity, temperature at which the conductivity was measured, activation energy, and structural prototype. Following this initial tabulation, the activation energy was excluded from the final database as it is not reported frequently enough to warrant inclusion.

Owing to the complexities described above, further expert validation of the data was required. The ionic conductivity of a material is typically determined using ACIS, although it can also be calculated through molecular dynamics simulations ²¹, or examined by NMR diffusion experiments ²², ion migration studies ²³, or entirely different measurements not directly related to ion transport (e.g. maximum entropy method analysis of diffraction data ²⁴). Even experimental papers which report a measured conductivity for a material through ACIS may themselves involve a variety of measurements and sample preparations, creating uncertainty around reported values. Postgraduate and postdoctoral researchers with >2 years direct experience of battery research with a broad knowledge of background literature assessed experimental procedures, consistency in sample preparation, quality, and other aspects of the reported data based on the details provided. Each researcher handled a selection of entries and was tasked with validating the database entry against the source report.

Dealing with such a large table of data in spreadsheet form adds significant challenges. Specifically, working with an online spreadsheet directly with twenty researchers leads to issues with version conflicts, edit histories, issues with concurrent user access, merging changes from multiple users, as well as assigning and tracking tasks. These issues were avoided by reducing the individual tasks to their core components through a bespoke interface developed with the streamlit prototyping library, shown in Figure 1. The interface was created to present a single entry from the database with its composition, associated conductivity at a specific temperature, and source paper. For each entry, the researcher was tasked with evaluating the conductivity at that specific temperature, making note of any mistakes with the composition and reported conductivity or temperature from the source. Positive feedback to researchers was provided through the presentation of a unique compliment provided by a GPT-2 transformer based language generation model ^{25,26}, displayed to the researcher after evaluating and recording each entry (Note 1).

Note 1: A sample of unique compliments that were created using a prompt that is constructed of a predefined complimentary phrase that is engineered to induce positive sentiment, followed by the researchers name. The remainder of the sentence is created by a GPT-2 language generation model.

- *Mendeleev smiles down on your work from above Cameron! Cameron: I do?*
- *The world would be a better place if there were more people like you Mike! The player to join the team. A name to remember. A reward for winning the game, to buy or*
- *On a scale of 1 to 10, you're at 11 Luke! The best method is the traditional 1:5 formula, which applies the weight of all of the other 1's when they are*
- *I have no mouth, but I must sing your praises Matthew! And now come, let us meet from heaven; there we'll sing, and come and sing as we go.*
- *You're a star Alex! Have you seen this or heard of me? My name is Alex Jones and I am a 24-hour Twitter personality and pundit*
- *You bring out the best chemist in other people Michael! I will always help you grow some more. I thank you all for choosing the Biodiesel Institute. Thanks again so much.*
- *In the robot uprising, I'll make sure you're given the best treatment Manel! The world was saved by everyone, and all the robots have escaped the prison and are on their way to rescue our hero*
- *You're a data wizard Matt! You do not need to work with large data sets – because you will save many minutes and dollars in learning about the big data.*

Which Researcher are you

Luke Daniels

Record Status

Pending

Select the current record to process

78

Password

Current Formula to Verify:

Li1.3Ti1.75Co.3P3O12

Parsed Formula:

Li0.071 Sc0.016 Ti0.093 P0.164 O0.656

Previous Comments

201108154538: Source paper was incorrect, this has been corrected. No data on this materials ionic conductivity was given in review. I have input it from new primary paper at 200 C

Tick boxes to view source and review paper

☐ Review Paper
 ☒ Source Paper

Source Paper

Recorded Conductivity, σ (S/cm): 2.30×10^{-5} at 200 °C

New Conductivity (leave blank if already correct)

Conductivity Comment

Temperature Comment

Any other things to know

Any additional comments

This is of the cooling quenched sample

Record Status

New Status (Make sure to delete any previous values)

Validated

Submit

Saved!

You're the hardest working hobbit since Samwise Gamgee Luke! All the reporters say this is the best you've ever done!

Made with Streamlit

Figure 1: A screenshot of the streamlit interface developed to present each researcher with a single database entry to validate the data.

Database Overview

A database was created with 820 entries collected from 214 sources; each entry contains the ionic conductivity of a chemical composition at a specific temperature, ranging from 5-873 °C, with an expert-assigned structural label. There are 434 different entries (Table 1) in the database for ionic conductivities experimentally measured at room temperature (15-35 °C). For a further 31 materials, the room temperature conductivities are extrapolated

from measurements above room temperature, to obtain a dataset of 465 entries, with 403 unique compositions, as 37 room temperature compositions have conductivities extracted from multiple reports.

Table 1: The number of experimentally reported conductivities contained within this dataset.

| Description | Count |
|---|--------------|
| ACIS measured conductivities at any temperature | 789 |
| ACIS measured conductivities at room temperature (15°C to 35°C) | 434 |
| Room temperature conductivities extrapolated from higher temperature | 31 |
| Total number of conductivities at room temperature | 465 |
| Total number of conductivities at any temperature | 820 |
| Number of unique compositions with a conductivity at any temperature | 455 |
| Number of unique compositions with a conductivity at room temperature | 403 |

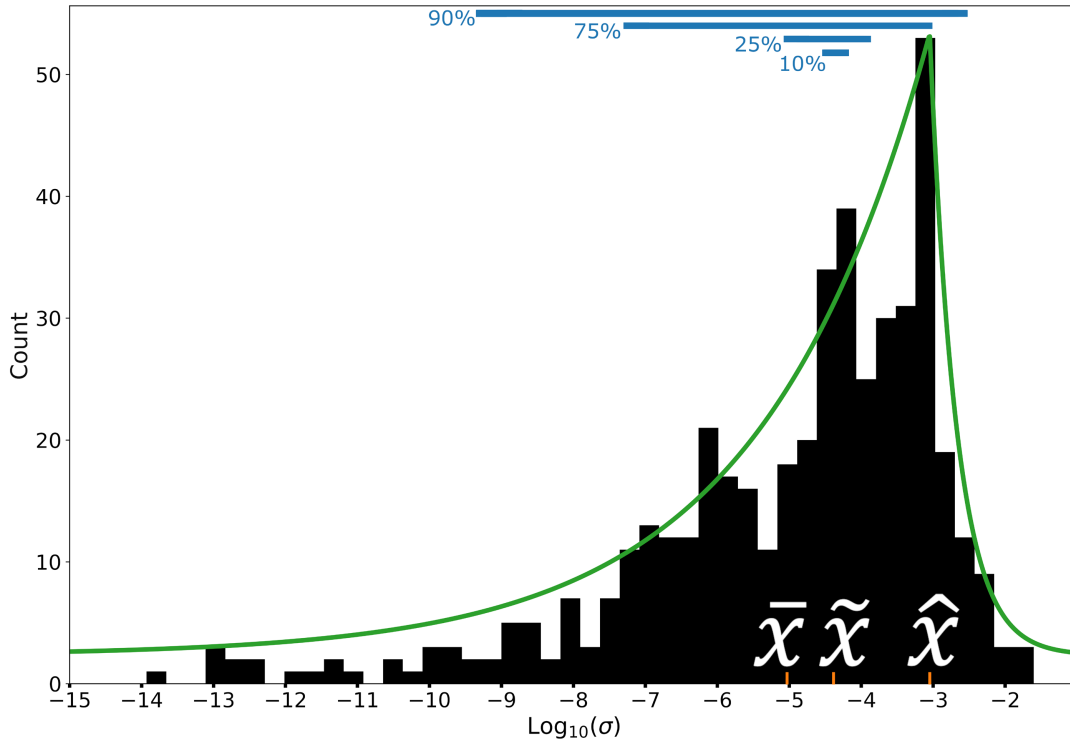


Figure 2: Distribution in room temperature conductivities for materials in the dataset. A histogram displaying the 465 room temperature conductivities (in units of $S\text{ cm}^{-1}$) from materials contained in this dataset and the relative distribution of their $\log_{10}(\sigma)$. The mean (\bar{x}) value of -5.01 , the median (\tilde{x}) value of -4.41 , and the mode (\hat{x}) value of -3.05 are marked on the x axis. An asymmetric Laplace distribution has been fit to this data, overlaid in green. The count of each bar is given on the y axis, with the percentage of materials falling within each percentile range around the median overlaid on the top axis.

The room temperature conductivities span the range of 5.00×10^{-16} $S\text{ cm}^{-1}$ to 2.50×10^{-2} $S\text{ cm}^{-1}$, with a mean $\log_{10}(\sigma)$ of -5.01 and median of -4.41 (Figure 2). The distribution of conductivities in this dataset and the associated standard deviation are estimated by optimizing the parameters of many probability distribution functions using the Fitter library (github.com/cokelaer/fitter); the distribution which fits the data with the lowest error is an asymmetric Laplace distribution. The interquartile range (50% of the data; materials from the 25th to the 75th centile of $\log_{10}(\sigma)$ in the dataset) spans from -7.30 to -3.03 , which is taken from the dataset values directly.

During database construction, each material in the dataset was manually allocated a label, based on the structural prototype the material belongs to. If the material structure was not discussed directly in the text and its family could not be deduced with reasoning, then this composition was assigned the structural label of Other. The breadth of structural chemistry encompassed by this dataset is shown by the fifteen unique families present in this set of expert-curated labels Table 2, which can be used to partition this database and expose trends that have been reported in the literature.

Table 2: Each of the 403 room temperature conductivities grouped by expert assigned structural labels with the count and simple statistics of the range of conductivities for each family.

| Structural Family | Count | Average $\log_{10}(\sigma)$ | Std. Dev. $\log_{10}(\sigma)$ | Max $\log_{10}(\sigma)$ | Min $\log_{10}(\sigma)$ |
|--------------------------|--------------|---|---|---|---|
| Anti-Perovskite | 8 | -6.07 | 1.76 | -3.03 | -8.14 |
| Argyrodite | 21 | -3.6 | 2.03 | -1.85 | -10 |
| Garnet | 50 | -4.24 | 1.21 | -2.69 | -8.06 |
| Glass | 31 | -6.16 | 1.98 | -2.23 | -8.89 |
| Glass-Ceramic | 5 | -5.72 | 1.27 | -4.9 | -7.95 |
| LISICON | 26 | -8.09 | 3.3 | -4.15 | -13.04 |
| Lysonite | 2 | -10.25 | 1.44 | -9.23 | -11.27 |
| NASICON | 132 | -4.72 | 1.45 | -2.21 | -10 |
| Olivine | 6 | -9.53 | 0.92 | -8.85 | -11.2 |
| Other | 17 | -5.68 | 3.05 | -2.8 | -15.3 |
| Perovskite | 58 | -4.29 | 1.25 | -2.8 | -8 |
| Phenakite | 3 | -10.7 | 0.24 | -10.5 | -10.97 |
| Rocksalt | 8 | -7.79 | 2.83 | -4.87 | -13.93 |
| Thio-LISICON | 32 | -4.05 | 1.64 | -1.6 | -7.9 |
| Zircon | 4 | -7.39 | 2.31 | -4.85 | -10.18 |

In Figure 3 the distribution of $\log_{10}(\sigma)$ for each structural family for which room temperature data is available, has been created by fitting a density kernel to the conductivities. This consists of placing a Gaussian distribution of fixed height and width at the x co-ordinate for each conductivity, and summing these together to approximate the probability density, allowing us to estimate the spread of reported conductivities. Irregular distributions with long tails are observed for some structural families. As the majority of these sets contain fewer than 50 reported materials, reports of materials with higher conductivities in the literature will lead to anthropogenically biased distributions 27.

Anthropogenic bias is inescapable when constructing a dataset of experimentally measured property from the literature. The reduced scientific interest in undertaking the lengthy characterization of materials with little importance to electrolyte chemistry, has meant that materials with very low or negligible conductivity are underreported. Distributions will be skewed towards conductivities of interest, and thus not truly representative of the underlying chemistry.

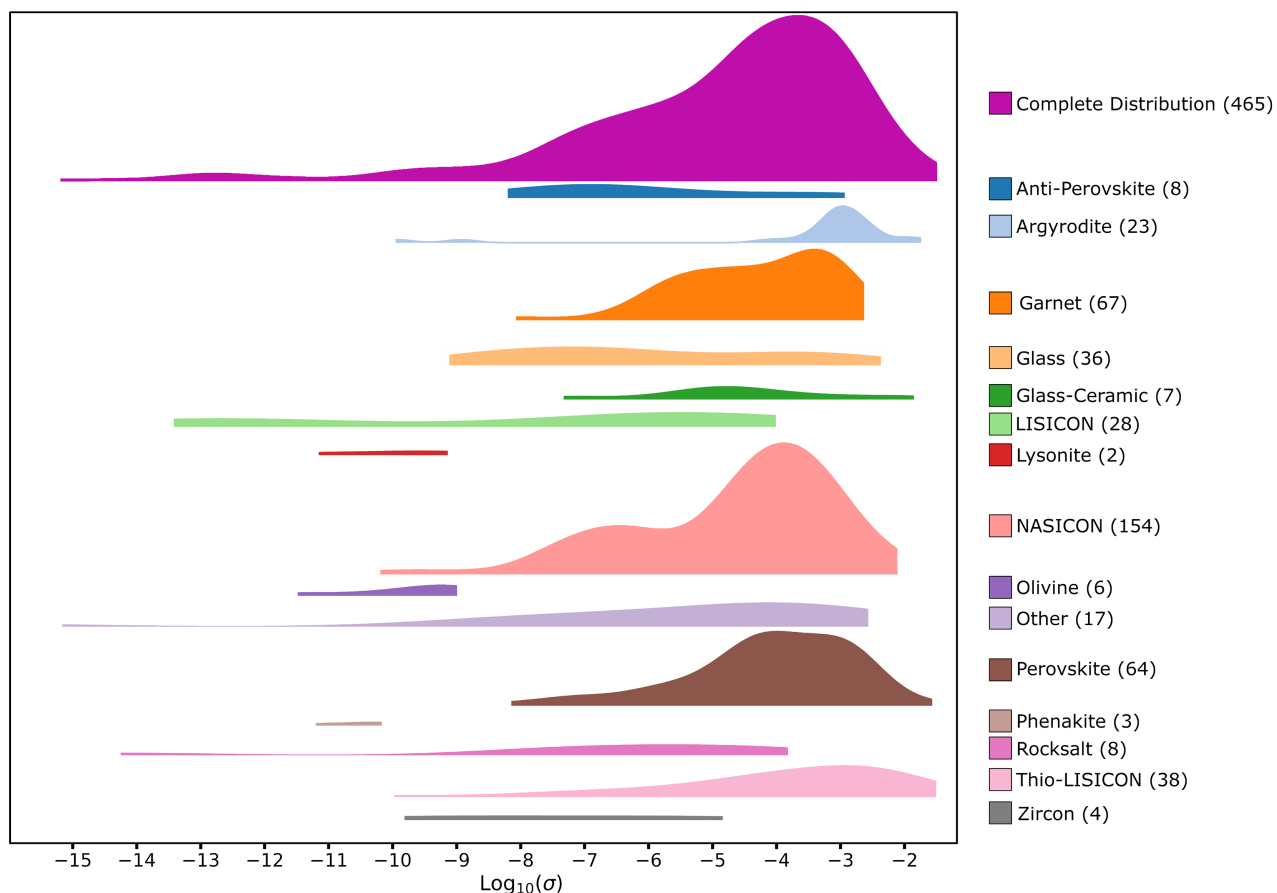


Figure 3: Distribution of room temperature conductivities across expert-curated structural families. Fitted distribution functions of the room temperature $\log_{10}(\sigma)$ for all materials within the database separated into expert-curated structural families and scaled by the number of entries within each family, given in brackets.

The room temperature dataset predominantly consists of NASICON, garnet, perovskite, glass, thio-LISICON, and LISICON type materials, each with more than 27 members. The anion chemistries of the materials are provided in Table 3, showing that 75% of the materials in the database are pure oxide compounds (consisting of 44% NASICON, 19% garnet, 18% perovskite and 8% LISICON type materials), 12% are pure sulphides, and 2% are pure halide compounds. Mixed anion materials (oxyhalides, oxysulphides, etc.) make up 11% of the materials included (46% of these are argyrodites such as $\text{Li}_6\text{PS}_5\text{Cl}$, and 16% are anti-perovskites such as Li_3OCl). In general, materials containing sulphur as an anion exhibit higher minimum and

maximum conductivities which is supportive of the outlook that is commonly encountered in the literature that sulphides exhibit the highest Li ion conductivities.

Table 3: Chemistries of the materials in the database of 465 room temperature Li ion conductivities based on anions: pure oxides, oxides containing at least one other anion, pure sulphides, sulphides containing at least one other anion, pure halides, and other (which contains materials such as LiBH_4 , Li_3P and $\text{Li}_2\text{Ca}(\text{NH})_2$). The minimum and maximum Li ion conductivities at room temperature are given for each group.

| Materials | No. entries | σ_{\min} (S cm^{-1}) | σ_{\max} (S cm^{-1}) |
|------------------------------|-------------|--|--|
| Oxides | 347 | 5.00×10^{-16} | 6.31×10^{-3} |
| Oxide with other anion(s) | 18 | 1.00×10^{-10} | 9.38×10^{-4} |
| Sulphides | 55 | 1.60×10^{-10} | 1.70×10^{-2} |
| Sulphide with other anion(s) | 32 | 8.13×10^{-9} | 2.50×10^{-2} |
| Halides | 7 | 1.18×10^{-14} | 1.51×10^{-6} |
| Other | 7 | 2.00×10^{-9} | 1.00×10^{-3} |

Machine Learning

With a database of materials gathered, unsupervised or supervised machine learning (ML) may be applied to these compositions to extract chemical trends. Unsupervised learning involves the application of embedding and clustering techniques based on the elements in the material, with no further knowledge of chemical properties such as conductivity required. Unsupervised techniques are beneficial as they do not require time-intensive labelling, and may highlight trends and similarities that may not be immediately apparent from a large collection of data in a table. Unsupervised clustering has successfully been applied in previous investigations

to cluster electrolyte materials ⁸ based on crystal structure through hierarchical clustering applied to the anionic frameworks of 528 Li containing structures from the ICSD. Conversely, supervised techniques attempt to fit a predictive function for a property to chemical descriptors such that the property can be predicted for a new material by statistical learning from known examples in a given training set. Machine learning is applied to compositional descriptors to predict each material's room temperature lithium ion conductivity (a regression task), or to predict whether each material possesses a room temperature lithium ion conductivity $\log_{10}(\sigma) > -4$ (a classification task).

In the previous chapter, we introduced the Element Movers Distance (ElMD) ²⁸ as a metric to quantify the similarity between two chemical formulae. This is demonstrated to be an expressive measure of chemical similarity that aligns with domain expert judgement. This metric can be incorporated with unsupervised dimensionality reduction and automated clustering to present chemical composition data to those who study these spaces. This brings high-dimensional compositional spaces into concise structured representations, such as maps, that can be interpreted by humans. In doing this the landscape of known compositions can be categorized according to our knowledge of related materials.

Following the methods described previously with the ElM2D plotting library (github.com/lrcfmd/ElM2D), we construct a distance matrix of ElMD scores between the compositions in the ICSD (2021) ²⁹ and the compositions contained within the ionic conductors database here. This metric space is reduced to two dimensions with principle component analysis (PCA) (Figure 4). A Gram centred matrix ³⁰ is first obtained from the given distance matrix, and then singular value decomposition of the Gram matrix carried forward to obtain the coordinates of each point projected to the first two principle components. PCA linearly scales each metric distance to maximally preserve each of the interpoint relationships across the dataset, which has previously been shown to closely

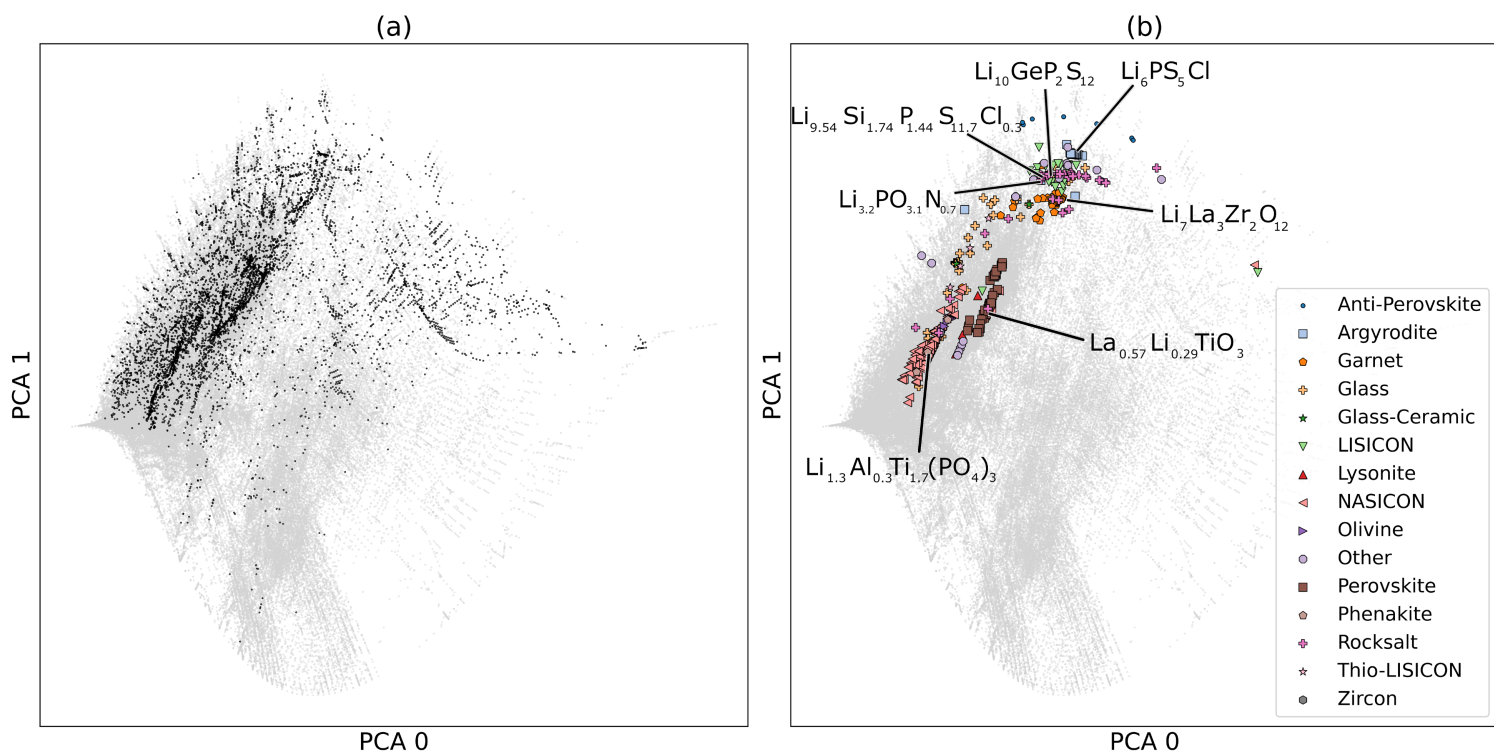


Figure 4: An embedding of the 127,638 unique compositions (grey) from the ICSD database (2021) with respect to EIMD similarity between compounds, embedded to 2 principle axes with PCA. (a) 6,972 of these compositions contain lithium (black), and (b) 455 unique compositions from this dataset with an experimentally measured conductivity at any temperature. The expert-curated structural label that each composition belongs to is indicated by the colour scheme given in the legend, with a selection of representative compositions and their embedded coordinates indicated.

reflect the true structure of the metric space²⁸. Figure 4 thus represents the distribution of this dataset in the compositional space of the materials constituting the ICSD.

Each of the lithium-containing compounds of the ICSD are highlighted against other compositions of the ICSD and the 455 unique compositions from our entire database (i.e., compositions with data recorded at any temperature) in Figure 4(a), with the expert-curated labels of the structural families included in the lithium conductors database in Figure 4(b). Though structure has not been included in the initial representation, expert-identified structural families are seen to tend to cluster in this compositional embedding, reflecting the connection between composition and structure. Perovskites (Figure 5(a)), NASICONs (Figure 5(b)), thio-LISICONs, and garnets are found in distinct areas of the

compositional map; each of these structural families are grouped tightly on the map, despite the absence of structural information (Figure 4(b)). The lithium ion conducting materials in the database are found in the same regions of compositional space as known lithium compounds, and can be seen to match the diversity of lithium chemistry that has been explored to date reasonably well.

This reflects the anthropogenic bias intrinsic to the research process, as much of the work devoted to discovering new lithium-containing materials has been driven by applications in battery technologies. There are a number of areas of accessible lithium-based chemistry (compounds seen on the right-hand side of Figure 4(a)) where known materials appear underexplored with regards to ionic conductivity. This compositional space should be considered in the search for new families of lithium ion conductors.

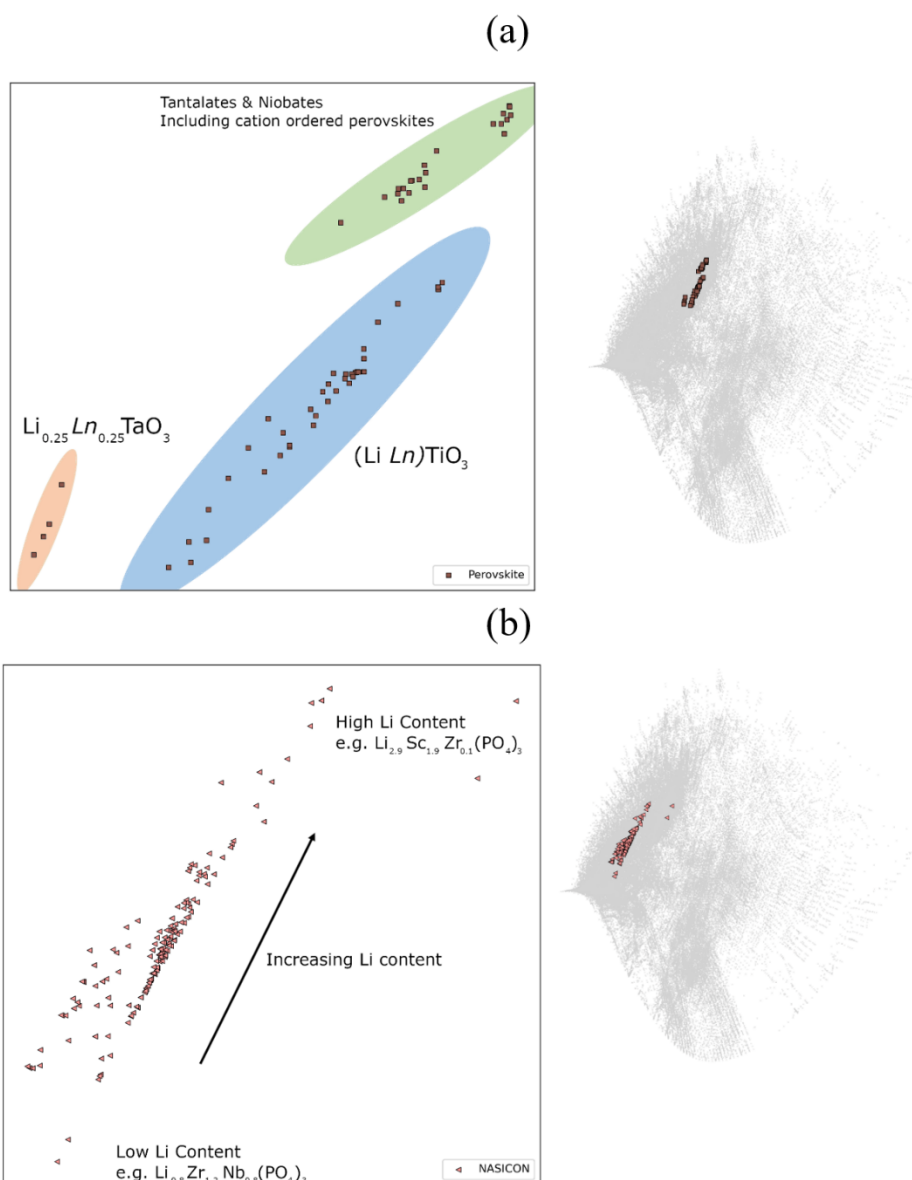


Figure 5: (a) The embedded positions of perovskite materials in the reported dataset when compared to compositions of the ICSD w.r.t. ElMD. Perovskites are split into three distinct regions, the most populated of which consists of $(\text{Li}, \text{Ln})\text{TiO}_3$ materials, with a small cluster of $\text{Li}_{0.25}\text{Ln}_{0.25}\text{TaO}_3$ compounds, and a final cluster dominated by tantalate- and niobate-based perovskites, also containing cation-ordered perovskites such as LiSrTiNbO_6 and $\text{LiSr}_2\text{Ti}_2\text{TaO}_9$. (b) The embedded positions of materials with NASICON structure type in the reported dataset when compared to compositions of the ICSD w.r.t. ElMD. The most populous structure type included in the database, assemble with clear dependence on Li content that increases from lower-left to upper-right of the cluster.

The previous chapter includes a discussion on PCA, remarking that whilst this gives an accurate realisation of compositional space with respect to ElMD²⁸, it is not the best representation for further processing with automated clustering techniques. The compact and concentric patterns that these clusters follow are difficult to unravel both visually and algorithmically, particularly when framed against the noise of so many unrelated compounds. We find that non-linear dimension reduction techniques attain a much clearer separation of the space into distinct regions of compositional similarity, which can be clustered more consistently (Figure 6). Uniform manifold approximation and projection (UMAP) draws apart the points of a space by first forming a neighbourhood graph of points in the metric space then embedding this graph to a 2-dimensional plane of projection via Laplacian Eigenmaps to capture global information³¹. These 2D distances are then refined through a ball and spring model³² to capture the local intricacies of the metric space.

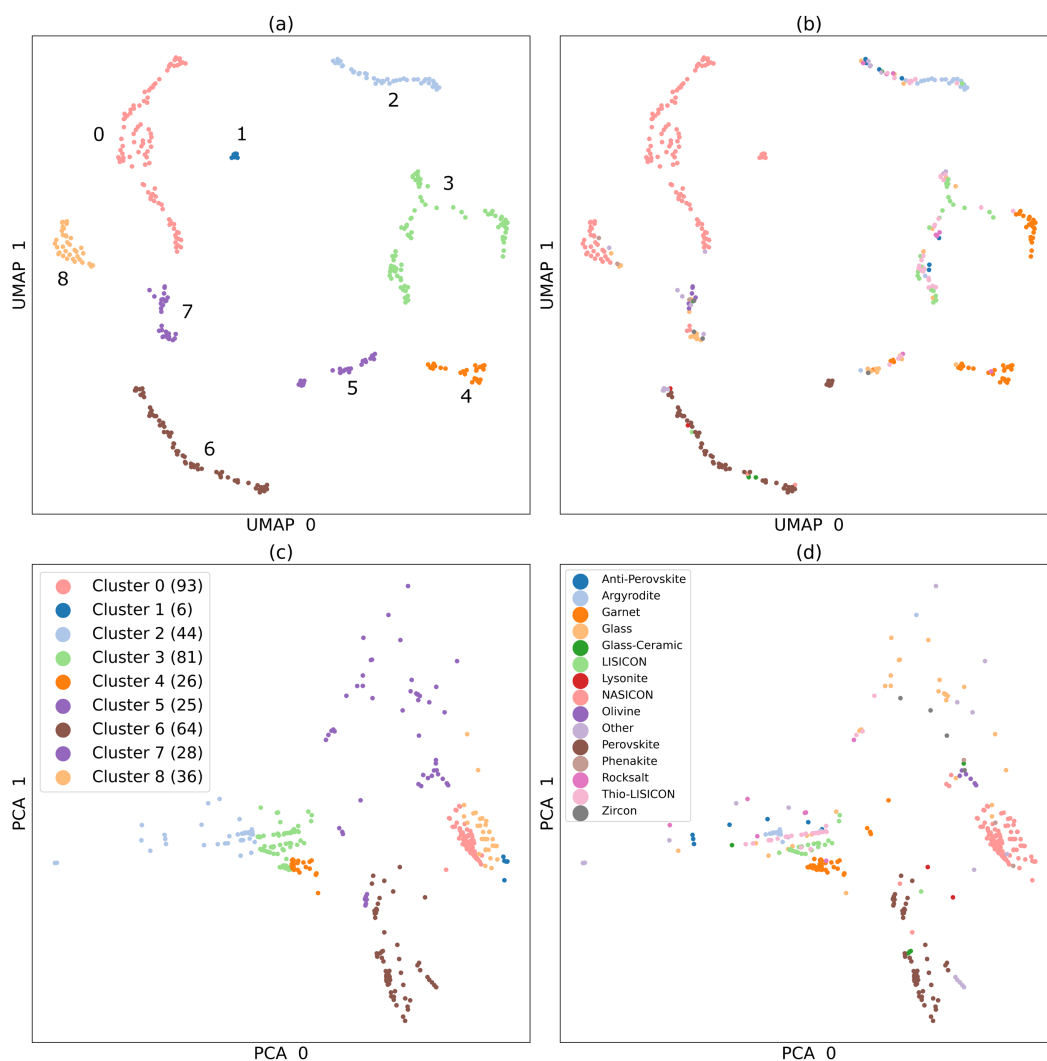


Figure 6: Embeddings of the 403 unique room temperature solid state electrolytes compositional data. (a) and (b) show the coordinates obtained from the UMAP embedding algorithm, whilst (c) and (d) arise from PCA. The cluster labels in (a) and (c) are obtained from the DBSCAN clustering algorithm applied to the UMAP embedded points in (a), with the number of materials in each cluster given in brackets. Cluster labels in (b) and (d) were assigned from expert review to classify each material under a structural prototype.

UMAP (Figure 6(a),(b)) and PCA (Figure 6(c),(d)) are applied to evaluate the reduced space of the 403 compositions of room temperature solid state lithium ion conductors in the database reported here. The UMAP plot contains several clear regions, which can be separated into 9 distinct clusters using the density-based spatial clustering of applications with noise (DBSCAN) algorithm³³ with an epsilon radius of 4 (Figure 6(a)). The epsilon value determines the radius of disks that are overlaid on every point in the 2-dimensional plot, which are then used to classify the points into different clusters. If two points cover each other with overlapping disks, then these will be assigned the same cluster label. DBSCAN has the ability to capture dense regions of an embedding, but if epsilon is too large then the output will fail to separate disjoint clusters. In this study, epsilon was chosen manually to maximise consistency between automated clusters and the clusters that can be visually observed.

Each of these unsupervised ML-derived clusters from Figure 4(a) are chemically reasonable, with clear stoichiometric substitutions or structural similarities connecting their constituents. This becomes apparent from comparison with the expert-derived structural family labelling in Figure 6(b) and 6(d). For example, Clusters 0 and 8 from the automated clustering are predominantly populated by NASICONs, perovskites are exclusively found in Clusters 5 and 6, whereas Cluster 4 is almost exclusively garnet structure materials. In addition to the practical benefits automated embedding and classification provides to rationally organise materials with minimal human bias, these clusters have further application in supervised training. As some data must be withheld from training and retained to test the performance of a trained

model, each DBSCAN-derived cluster will be used as a testing set in a process referred to as Leave One Cluster Out Cross Validation (LOCO-CV). These clusters range in size from 6 materials to 93 materials, with the training set then typically containing 85-90% of the available data to train each model. The distributions of $\log_{10}(\sigma)$ for each LOCO cluster have been plotted in Figure 7, with basic statistics given in Table 3, where many of the clusters span similar ranges of conductivity. Given the intra-cluster chemical consistency and inter-cluster dissimilarity, these assessments are a measure of how each model performs at predicting the ionic conductivities of materials that are chemically dissimilar from those on which the model has been trained.

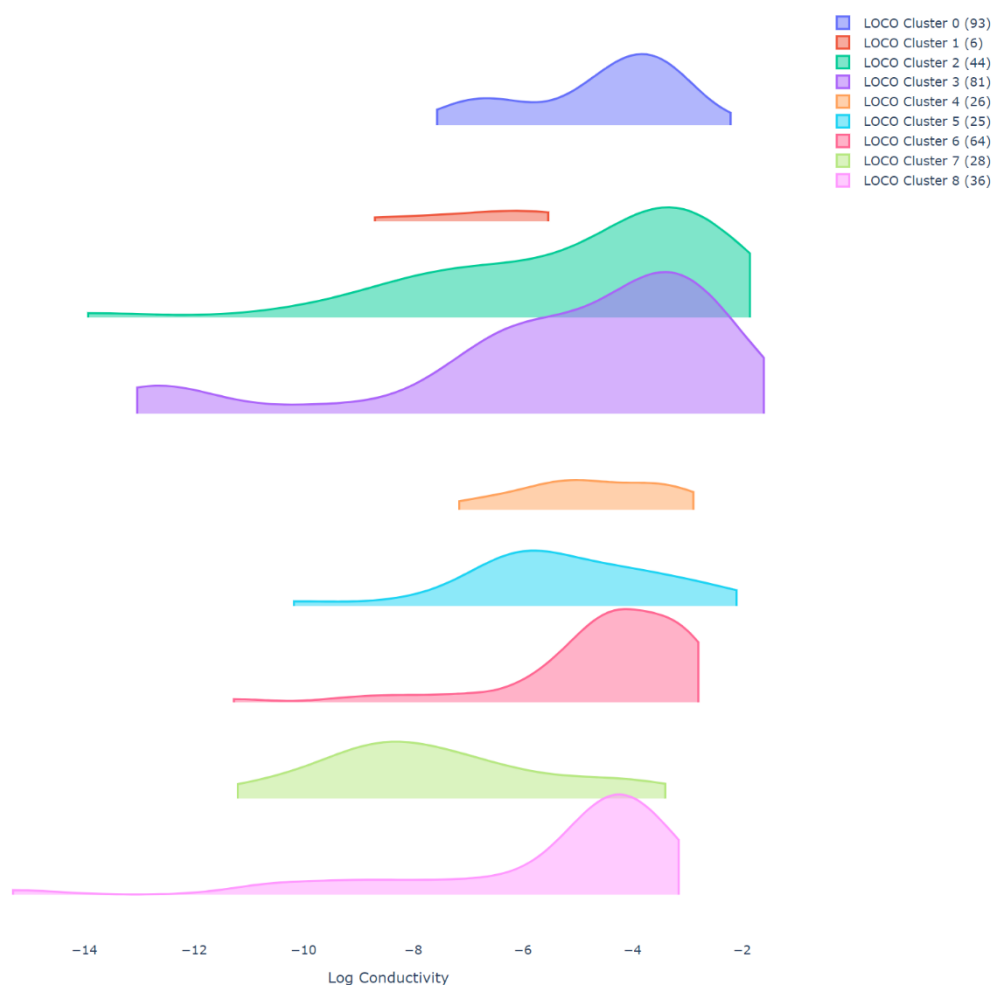


Figure 7: Distribution of room temperature $\log_{10}(\sigma)$ conductivities across each LOCO cluster from fitting a Gaussian kernel function, with the number of entries within each cluster given in brackets.

Table 3: Each of the LOCO clusters with the count and simple statistics of the range of conductivities for each cluster.

| LOCO Cluster | Count | Average $\log_{10}(\sigma)$ | Std. Dev. $\log_{10}(\sigma)$ | Max $\log_{10}(\sigma)$ | Min $\log_{10}(\sigma)$ |
|-------------------------|--------------|---|---|---|---|
| 0 | 93 | -4.58 | 1.38 | -2.21 | -7.56 |
| 1 | 6 | -6.7 | 1.2 | -5.54 | -8.7 |
| 2 | 44 | -4.91 | 2.5 | -1.85 | -13.93 |
| 3 | 81 | -5.28 | 2.98 | -1.6 | -13.04 |
| 4 | 26 | -4.74 | 1.22 | -2.89 | -7.16 |
| 5 | 25 | -5.28 | 1.8 | -2.1 | -10.18 |
| 6 | 64 | -4.49 | 1.61 | -2.8 | -11.27 |
| 7 | 28 | -7.59 | 2.07 | -3.4 | -11.2 |
| 8 | 36 | -5.54 | 2.63 | -3.15 | -15.3 |

Supervised Learning

A dataset of 403 entries is constructed, where compositions with duplicate room temperature conductivities from differing sources have been represented by the median of these multiple reported conductivities. With this dataset in hand, we apply the best available ML models that can be implemented with minimal modification, i.e. off the shelf. This is done with traditional statistical learners (ensemble models) with `mat2vec`¹⁴ composition-based feature vectors³⁴, and deep learning techniques (CrabNet). For statistical learners, we wish to ensure the best models and associated hyperparameters are chosen, so that we do not simply overfit to one portion of the data. A simple model with fixed hyperparameters is not guaranteed to give good predictions on unseen compounds. Such models may overfit to the training data, leading to poor predictions on unseen compositions, or give exceptional performance on certain subsets of the data with poor

performance on the rest. Some of the issues of overprediction can be remedied by surveying a range of statistical models ³⁵.

State of the art techniques for predicting materials properties through composition apply this principle by training an ensemble of models, in the belief that each model will learn to focus on a different set of features. The predictions of each individual model are combined, which tends to give more robust predictions across the entire domain. In statistical models, the ensemble approach is notably used in the random forest (RF) algorithm ³⁶, where large ensembles of decision trees are randomly constructed and kept or discarded depending on their predictive quality.

The resulting quality of RF predictions depends on the values of each hyperparameter chosen when initialising the model, and poor choices can lead to very poor models. To alleviate this, best practice has traditionally focussed on trialling a range of hyperparameters in combination with one another, but this is time consuming and does not guarantee that the optimal configuration will be found. More recent AutoML approaches ³⁷ improve on this by framing the choice of statistical model and its associated hyperparameters as a meta-problem to be solved. Many separate algorithms and hyperparameters can be trialled and assessed in combination, with the measured performance used to update a selection policy for future trials until optimal combinations are found.

In AutoSklearn ³⁸, many types of models and data pre-processing stages from the scikit-learn library are chained together to form data processing pipelines. The supplied training data is shuffled into k -folds cross-validation sets and used to assess each pipeline, with the performance noted. This performance is used to update the parameters of a tree-based Bayesian optimisation selection policy, which will decide the models and hyperparameters to choose in future iterations, alternating between exploring untried combinations, and exploiting relationships known to give good

results. Given that RFs return more robust predictions through ensembling many weaker models together, we would expect an ensemble of effective models to give even stronger predictions. As simple models are quick to train, thousands of pipelines can be evaluated during the AutoSklearn training process. After the allotted training time of ten minutes, the 50 pipelines with the highest performance are selected to form a trained ensemble which can be used to predict unseen data.

In comparison, Compositionally Restricted Attention Based Networks (CrabNets)³⁹ are an implementation of the transformer model⁴⁰ of deep learning. Here, self-attention is employed to learn how relationships between each of the elemental vectors in a composition are aligned with a target property. Each elemental vector representation is transformed by a trainable neural network layer, with the cosine similarity scores taken between each elemental latent representation. These scores are used to weight the contribution of each element before it is input into a fully connected neural network to predict the specific property. These elemental relationships improve as the model learns through back-propagation during training.

In this architecture, the transformers positional encoder is repurposed as a fractional encoder to capture the ratio of each element in the composition, which enables CrabNets to capture similarities and small variations in stoichiometry with precision. This is particularly relevant for ionic conductors, where minor substituents (e.g., those controlling the exact Li^+ content) can significantly influence the ionic conductivity because they determine the defect concentrations and associated local structure that can govern ionic motion.

One shortcoming of deep neural networks such as CrabNets is that they require large quantities of training data which are typically unavailable for materials science problems. This limitation can be alleviated by transfer learning, which involves pretraining

networks on much larger datasets of compounds and their associated properties, such as the computed energy of formation. The trained parameters of this network can be exported to initialise future models for different properties, as opposed to initialising all of these values randomly.

The desired benefit of pretraining the network on a wider range of compositions and their associated formation energies, is that the knowledge of chemical relationships absent in our training set can be extrapolated to future predictions. By transferring this knowledge from another domain, the most salient chemical relations are intended to be well represented in the network. This typically leads to a faster convergence to the optimal value when training the neural network on the desired property, and can lead to improved predictive performance in the target domain. This has been demonstrated in other investigations ^{41,42}, where the application of transfer learning and neural networks has achieved state of the art for materials property prediction. In this work we compare the performance of AutoSklern ensembles, randomly initialized CrabNets, and CrabNets that have been pretrained on compositions and their formation energies from the OQMD ⁴³.

Training CrabNets involves iteratively updating many model parameters of the network on the same dataset multiple times; each iteration is called a training epoch. Once an iteration has completed, the millions of model parameters will have been more finely tuned to align the data with the target property, which should give a better model than the previous iteration. When model training begins, we expect poor performance when predicting properties of materials in the test set, but as the model is further biased by training data after several epochs, more robust predictions should be attained. In general, when training neural networks, the training error steadily decreases over time, as the parameters of the model get more aligned with the input. After prolonged training however, these parameters begin to overfit to

the training data, and the model gets steadily worse at predicting anything outside the training set ⁴⁴.

The training and testing performance at each epoch can be plotted on a training curve, which characterises how performance evolves with the number of training epochs. A training curve can be used to determine the optimal training time (e.g. number of epochs). Model parameters can be exported from the training epoch that displays best performance at test set predictions. Training for sufficiently long time (to see degradation in test set performance) and then reverting to an earlier state in training is referred to as early stopping, in contrast to *a priori* deciding the number of training epochs, or training indefinitely. Early stopping across 500 training epochs is applied in this study, with each model taking the optimal set of training weights, giving a reasonable measure of how CrabNets with and without transfer learning perform using standard hyperparameters.

The performance of AutoSklearn and CrabNet regression and classification models at predicting the conductivities of the materials in this dataset is evaluated through four methods: Control studies, parity plots, scoring metrics, and cross-validation techniques. We then use the best approach from this assessment to train final regression and classification models on all available data.

To give some measure of the worst-case performance, we provide two control experiments. In the first control experiment, we take the reported conductivity of each material, shuffle these labels, and treat the average of five of these shuffled values as an ensemble prediction from a poor “model”. This has the effect of providing a quasi-random prediction that demonstrates how ensembles can bring predictions closer to the mean (Figure 8(a)). In the second control experiment, we demonstrate how a model which simply predicts the mean will perform. We take the mean of all of the room temperature conductivities (-5.02 in $\log_{10}(\sigma)$) and treat these

as the output prediction for each material, giving the same prediction for every entry. The true conductivities are plotted against each of these control predictions to observe the performance (Figure 8(b)).

Plots are an effective method to directly confirm the performance of a statistical model. For regression tasks, we plot the actual conductivities of each material against the predicted conductivities

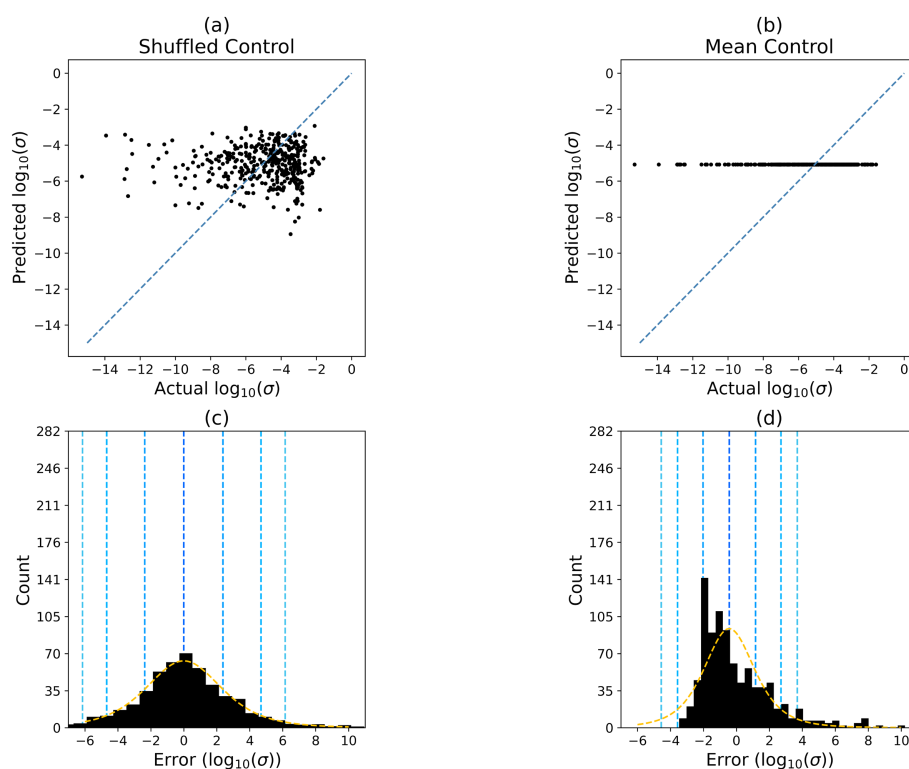


Figure 8: Parity plots and error distribution for two control studies. (a) The shuffled control parity plot demonstrates each materials actual conductivity plotted against an average of five randomly selected values across the dataset. (c) The distribution of errors across all experiments (without averaging) demonstrates the maximal error bounds we would expect from a poor statistical model, with 68% of predictions falling between -2.36 to 2.31 away from the true values. (b) The mean control experiment demonstrates the expected predictions for a model which has simply learnt the mean value of the dataset. Correspondingly, the distribution of errors (d) is simply a reflection of the distribution of conductivities around the mean value, and models which form predictions close to the mean will resemble this distribution. A Student's t-distribution (orange) is fit to the underlying data, with the mean of this distribution (dark blue), and the first, second, and third standard deviations away from this mean (light blue) overlaid in (c) and (d). A good model should have a mean of zero, with tight error bounds.

of a trained model. An ideal model would give each prediction perfectly on the leading diagonal. Dense pointclouds can be difficult to visually interpret, so errors of each prediction ($y_{\text{pred}} - y_{\text{true}}$) are calculated and plotted via histogram to quantify this distribution of errors. A Student's t-distribution is fitted to the errors of all repetitions (without averaging) to provide intervals for how many predictions are within certain bounds of error for each model. The shuffled control has a zero-centred gaussian distribution of errors on the histogram with a standard deviation of 2.34 (Figure 8(c)). The mean control has an error of -0.44 below the true value on average, with 68% of the predictions having an error within -1.99 to 1.10 of the true $\log_{10}(\sigma)$ (Figure 8(d)). Given this worst-case performance, we may demonstrate how the best compositional models perform at predicting new compositions.

When we have many plots for different models, it becomes difficult to visually confirm the “best” performing model. To quantify which of these models are best performing, we must use statistical metrics to rank the quality of the output predictions for each model. Regression models are often scored via Mean Absolute Error (MAE) and Pearsons R^2 score. The MAE returns the average difference between each prediction and its known value, where values closer to 0 reflect stronger model performance. The R^2 score shows the correlation between the true and predicted values, where a 1 is a perfect score, and anything below zero indicates that on average model predictions perform worse than simply returning the mean of the test set for all inputs.

For classification tasks, the performance may be demonstrated via a confusion matrix. This is a 2×2 matrix that compares the predictions made by the classification model against the true classification labels. An ideal result would have leading values (True Positives and True Negatives) and zeros elsewhere, but in reality many predictions will be False Positives and False Negatives. For simplicity however, the most frequently reported score for classification is accuracy. The accuracy score is defined

as the number of true predictions divided by the total count of values in the testing set:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{Eq. 1})$$

On heavily imbalanced datasets with few negative class instances, the accuracy can return a high score for poor classifiers that output a single classification. This is due to the small number of negative instances, which do not significantly alter the denominator even if they are heavily misclassified (Eq. 1). To prevent misleading reporting, the Matthews correlation coefficient (MCC)⁴⁵ can be taken as a more informative score⁴⁶ by considering the proportion of each class in the confusion matrix:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (\text{Eq. 2})$$

The MCC is calculated by taking the difference of the product of true predictions and the product of false predictions, and dividing by the geometric mean of all entries in the confusion matrix. This returns a value from 1 for perfect classifications to -1 for entirely incorrect classifications. The MCC provides more weighting to the score for any misclassified values, allowing us to judge the outcome of the confusion matrix succinctly. By themselves, isolated scores do not convey the strength of a model and these must be compared against a known point of reference, such as a control study, to understand the significance of a particular result.

As an aim of machine learning models is to predict the behaviour of as-yet unknown materials, it is important to distinguish between performance in interpolation between materials that have similar chemistries, where similar structure-property-composition relationships would be expected, and in extrapolation to materials characterised by structure and bonding that is not found in the training set. For example, predicting performance within a solid solution family with some members in the training set used would

be interpolation, whereas evaluating the conductivity from a material with a new structure type would be extrapolation.

This question naturally arises when evaluating ML model performance. Here, it is important that the data being tested have not been previously used to train the model, but in and of itself, this does not directly address interpolation versus extrapolation ability. The standard method of splitting data is via k -folds cross-validation, where the dataset is split into k equal sets, and one of these sets is used to test the model. In this report we take $k = 5$, where the model is trained on four of these subsets (80% of the data) and then tested on the fifth (20% of the data). This process is repeated for each set, and the mean score across all test sets is used as the final measure of performance. As many of the compounds in this dataset possess some similarity with one another, we expect the model should be able to interpolate relationships between known compositions.

Ideally, we want predictive models to be able to extrapolate beyond known materials, and statistically infer future chemical relationships from observed compositions. To test this, we utilise the DBSCAN labels assigned in Figure 4 as Leave One Cluster Out (LOCO) labels to separate the 403 unique room temperature conductors into testing sets. As the compositions within each cluster have been confirmed to share chemical similarity, and to have dissimilarity from other clusters, using each cluster shown in Figure 4(a) as a testing set provides a better estimate of the ability of a model to screen novel compositions than the k -folds approach, which will entail greater chemical similarity between the training and testing sets. Time split holdout is a common technique, where all materials that have been discovered before a certain date are used for training, with those discovered after this date used for testing. As time of discovery was not recorded in our dataset, this was not used to judge models, however an experimental holdout set of more recently discovered materials is used to test the final models in the next section.

Both of the applied cross-validation techniques are used to train AutoSklearn and CrabNet regressors and classifiers, with the average of five repetitions of each experiment taken as the final score. We collate the performance of the two control studies and the ML models for regression and classification, in Table 4 and Table 5 respectively.

Table 4: Regression Performance Metrics, average results of each regression model, judged by Mean Absolute Error and Pearsons R^2 metric under both dataset cross-validation regimes. The average value of the training performance across the test sets is first calculated for each metric, and then averaged across each of the five repetitions; standard deviation shown in brackets.

| Model | MAE (k -folds) | R^2 (k -folds) | MAE (LOCO) | R^2 (LOCO) |
|----------------------|--------------------|---------------------|--------------------|--------------------|
| Shuffled Control | 2.31 (0.06) | -0.99 (0.13) | 2.43 (0.08) | -2.6 (0.4) |
| Study | | | | |
| Mean Control Study | 1.71 (0.0) | 0 (0) | 1.72 (0) | -0.46 (0) |
| AutoSklearn 2.0 | 1.10 (0.04) | 0.46 (0.05) | 1.62 (0.08) | -0.4 (0.2) |
| Randomly Initialized | 0.96 (0.02) | 0.55 (0.03) | 1.131 (0.006) | 0.15 (0.03) |
| CrabNet | | | | |
| Transfer CrabNet | 0.85 (0.02) | 0.62 (0.02) | 0.99 (0.03) | 0.33 (0.02) |

The two control studies give the highest MAE and lowest R^2 scores between the actual and the predicted values under each cross-validation scheme. These numbers are important to consider when evaluating any improvement in predictive performance. All models perform better than these controls, and under k -folds cross-validation, and AutoSklearn models perform comparably to randomly initialised CrabNet models. However, under LOCO-CV, the AutoSklearn model fails to fit a suitable decision boundary to predict unseen materials; performance metrics reveal no significant improvement over the mean control. CrabNet models are better than AutoSklearn models at the extrapolatory LOCO task, and

these see improved performance in both MAE and R^2 correlation. CrabNet models with transfer learning outperform all other models across each metric and cross-validation scheme. The $\sim 10\%$ increase in performance of transfer learning regression models over those initialised randomly suggests that pretraining in other domains has given the model a clear advantage when inferring unseen chemical relationships. To demonstrate this further, three of the regression models parity plots and distribution of errors are given in Figure 9. These plots allow us to visually judge models against one another, and to assess each model’s performance at predicting materials similar to those within the training dataset (k -folds) as opposed to materials with unseen chemistry (LOCO-CV).

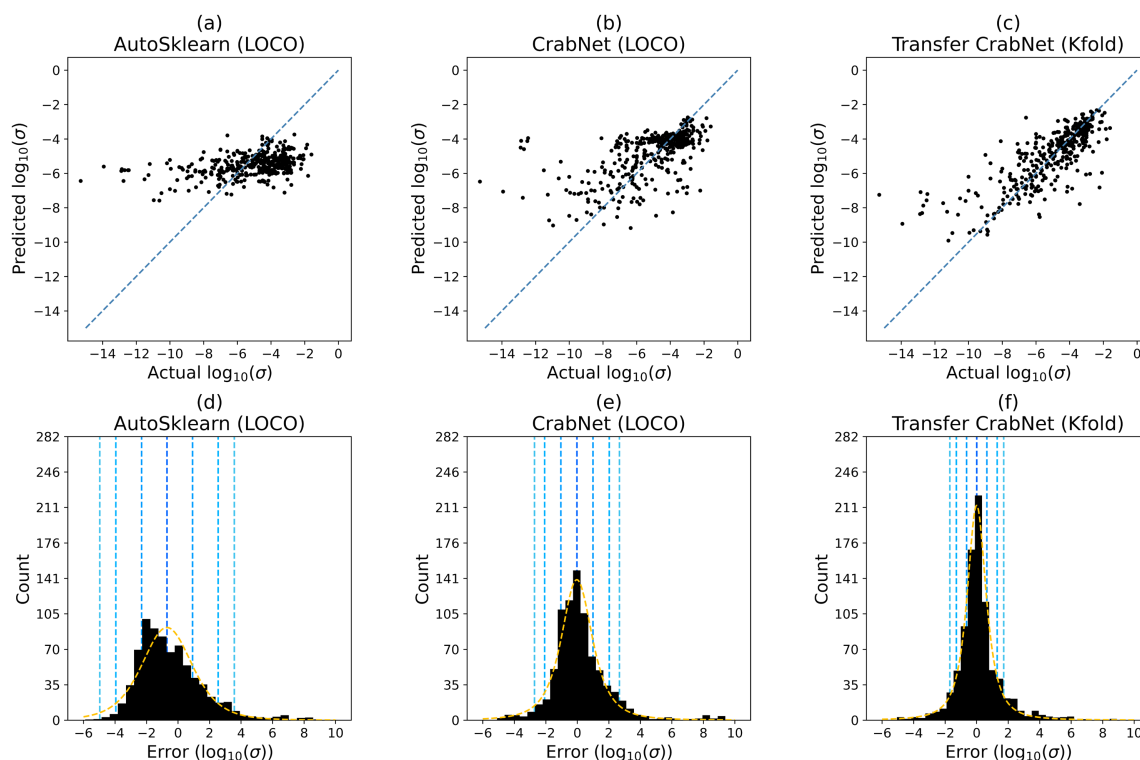


Figure 9: Parity plots and error distributions for three regression models. AutoSklern models assessed under LOCO-CV (a), (d) share the most similarity to the controls in Figure 5, and are thus judged to be the least effective ML model under investigation. Under LOCO-CV, CrabNet models with transfer learning offer improved performance, which can be visually confirmed by the spread of points falling closer to the leading diagonal (b), and the distribution of errors being centred around 0 with a smaller standard deviation. ML models give a tighter distribution of errors when validated with k -folds, with transfer learned CrabNets possessing the most favourable actual vs. true characteristics (c) and distribution of errors (f).

The AutoSklearn regression model under LOCO-CV (Figure 9(a)) demonstrates tighter prediction error bounds than the shuffled control, but still leads to predictions with an error of -0.68 on average and a standard deviation of 1.55 (Figure 9(d)). An ML model which typically achieves predictions of ionic conductivity within two orders of magnitude could be interpreted as a positive outcome. However, comparison to the mean control demonstrates that this model has not learned a meaningful representation for extrapolating beyond the chemistries within the training set. The AutoSklearn error distribution is not an improvement over the mean control, which has an average error of -0.44 and a standard deviation of 1.54 (Figure 8(d)). CrabNets with and without transfer initialization output a range of predictions closer to the real values, with tighter error bounds than AutoSklearn models.

The CrabNet regression models with transfer learning trained under LOCO-CV (Figure 9(b)) are not as consistently skewed as AutoSklearn, with an average error of -0.02 and a standard deviation of 0.811 (Figure 9(e)). These models typically return predictions with less error for high and medium conductivity materials, but often fail to capture the outlying low conductivity regions. This highlights the complexity of predicting exact materials properties when there has been little exposure to these unexplored chemistries. The best regression performance is achieved using CrabNet models with transfer learning under k -folds cross-validation (Figure 9(c)), which leads to a distribution of errors centred around -0.01, and a standard deviation of 0.58 (Figure 6(f)). As LOCO-CV forces each model to extrapolate future predictions, it is expected that the figures of merit will be less attractive than under k -folds cross-validation. Whereas regression models achieve only a modest improvement to the bounds set by the respective control studies, this is not the case for each of the classification models, which we turn to now.

Table 5: Classification Performance Metrics Average results of each classification model predicting whether materials possess $\log_{10}(\sigma) > -4$, judged by Matthews Correlation Coefficient (MCC) and accuracy under both dataset cross-validation regimes. The average value of the training performance across the test sets is first calculated for each metric, and then averaged across each of the five repetitions; standard deviation shown in brackets.

| Model | MCC (<i>k</i>-folds) | Accuracy (<i>k</i>-folds) | MCC (LOCO) | Accuracy (LOCO) |
|----------------------|-----------------------------|----------------------------------|--------------------|------------------------|
| Shuffled Control | -0.02 (0.03) | 0.50 (0.02) | 0.00 (0.07) | 0.52 (0.03) |
| Study | | | | |
| Mean Control Study | 0 (0) | 0.58 (0) | 0 (0) | 0.64 (0) |
| AutoSklern 2.0 | 0.46 (0.04) | 0.74 (0.01) | 0.10 (0.05) | 0.63 (0.03) |
| Randomly Initialized | 0.57 (0.01) | 0.786 (0.006) | 0.36 (0.01) | 0.62 (0.01) |
| CrabNet | | | | |
| Transfer CrabNet | 0.633 (0.002) | 0.814 (0.009) | 0.38 (0.01) | 0.71 (0.01) |

Table 5 displays the average MCC and accuracy score for each models test set performance across five runs, where it is seen that control models may seem initially reasonable when judged by accuracy. CrabNet models with transfer learning return the highest MCC of 0.63 under *k*-folds cross-validation, and CrabNets without transfer learning return a slightly lower score of 0.57. AutoSklern models do not give as strong performance, with an MCC of 0.46, but this is clearly a step improvement on the MCC scores of the control studies, with accuracy also seen to improve by some margin when comparing each model to the controls. As with the regression models, classification models trained under LOCO-CV return lower scores. This is highlighted by the AutoSklern model, which has a particularly poor MCC (close to the MCC of zero of the two controls) of 0.10 when classifying LOCO test set materials, despite a promising accuracy score. The highest scoring LOCO

classification model is the CrabNet with transfer learning; an MCC of 0.38 indicates more of the high conductive materials are correctly classified as having $\log_{10}(\sigma) > -4$ than misclassified, which is supported by the high test set accuracy of 0.73.

The two distinct cross-validation techniques have been applied to rank these statistical models against one another. However, interpolation between related materials within known chemistries (defined as known structure and bonding) should be considered independently from extrapolating into unknown chemistries beyond the training data. Accordingly direct comparison should not be drawn between the metrics for the two different cross-validation protocols, as these assess different aspects of the performance of the ML models trained against the dataset. We are forced to use the data in our possession to assess the quality of each model. The data arise from the efforts of researchers in the field, and thus reflect various research trends and foci that have emerged, rather than directly expressing the possibilities for structure, bonding and performance for materials drawn from element combination at the level of the periodic table. Given this anthropogenic bias, there will be consistencies and trends within each chemical family of the dataset.

By separating the materials of the database into clusters by chemical similarity and testing under LOCO-CV, the reduced performance compared to validation by k -folds highlights the challenge of extrapolating known compositional relationships to other chemical families that may span different ranges of conductivity. Comparatively, under k -folds cross-validation, each material in the testing set has a greater likelihood of having corresponding materials with similar elemental composition to their own in the training set. The model under assessment thus has more opportunities to interpolate between compositions in the training data, allowing it to make stronger predictions as it has to some extent been presented with similar examples during the training, rather than having them deliberately withheld.

This emphasises the strength of structure-property-composition relationships in lithium ion transport. It is reasonable to assume that ion transport takes place by local hopping through barriers governed by physical models that are closely connected in their physiochemical origin across all materials in the dataset regardless of structure and bonding. However, the changes in structure and bonding between these machine-identified materials clusters in which lithium transport occurs by similar, unifying diffusion mechanisms are sufficient to hinder extrapolation of performance from one set of chemistry to another, despite no fundamental change in mechanism taking place between the clusters.

This contrasts with the situation prevailing for example in superconductivity, where entirely different mechanisms may govern high-temperature superconductivity in cuprates and low temperature superconductivity in elemental and alloy systems that pair by weak-coupling BCS. This mechanistic difference has been shown to undermine attempts to extrapolate with machine learning from superconductors with one pairing mechanism to another⁴⁷, whereas for lithium ion transport it is the chemistry (the structure and bonding) that controls performance even under a unified physical mechanism. Nevertheless, CrabNet models with transfer learning are seen to consistently outperform both the control studies and AutoSklearn models at predicting ionic conductivity. This is shown statistically across all cross-validation schemes and metrics in both classification and regression models, and can be visually attested from the parity plots. As such further discussion will assume these models as the focus unless stated otherwise.

The Final Models

When screening compositions with machine learning we want to use the best possible model to increase the likelihood of making robust predictions. Model performance is typically improved by using the most training data available, and choosing an optimal

training time. As discussed earlier, the optimal training time can be determined by assessing the performance vs. epoch training curve to decide which set of model parameters to use (i.e. early stopping). An important practical consideration is that any model to predict ionic conductivity would be most valuable when screening new materials. Accordingly, to assess the ability of our ML models to estimate the ionic conductivities of unstudied materials or novel chemistries, we train a final classifier and a final regressor on the entire initial database of unique room temperature conductivities and test it against eleven newly reported materials that have not been included in the initial database. We refer to this new set of materials as the experimental holdout set. These are selected to represent a range of chemistries and also conductivities, which matches the situation facing the experimentalist targeting new families of ion-transporting materials: it is desirable to understand the likely lithium conductivity of a particular composition in order to aid the selection of specific new chemistries for investigation.

We select CrabNet with transfer learning as the architecture for these two models, as k -folds and LOCO-CV assessment show that it offers the best interpolation and extrapolation performance based on the considerations above. The final CrabNet models are trained on all unique entries of the initial database presented here. In the earlier validation investigations, early stopping could be employed by using the test data to select the set of network weights at the best performing training epoch on the training curve. In our final models, a fixed number of training epochs are determined *a priori* by assessing the training curves of CrabNets with transfer learning under LOCO-CV and selecting a training time which typically attains optimal performance. Final models are trained on all unique compositions with room temperature conductivity (*i.e.*, all 9 LOCO clusters), with the classification model trained for 98 epochs, and the regression model trained for 323 epochs.

The performance of these neural networks at classifying or predicting the $\log_{10}(\sigma)$ of a selection of recently reported materials is assessed across a range of reported conductivities. The individual performance for each material in the holdout set is given in Table 6.

Table 6: Final regression and classification model predictions of the experimental holdout set. CrabNets with transfer learning are trained on all 403 unique compositions and the associated $\log_{10}(\sigma)$ or classification target at room temperature. The experimentally measured $\log_{10}(\sigma)$ of each of the 11 materials in the holdout set are given alongside a predicted $\log_{10}(\sigma)$ and conductivity class for the material from the final models, the boundary against which the classification is performed has been marked in black.

| Composition | Measured Conductivity ($\log_{10}(\sigma)$) | CrabNet Regression Prediction ($\log_{10}(\sigma)$) | CrabNet Classifier Prediction ($\log_{10}(\sigma) > -4$) |
|---|--|--|---|
| $\text{Li}_{10.35}\text{Ge}_{1.35}\text{P}_{1.65}\text{S}_{12}$ ⁴⁷ | -1.85 | -3.60 | 1 |
| $\text{Li}_{10.35}[\text{Sn}_{0.27}\text{Si}_{1.08}]\text{P}_{1.65}\text{S}_{12}$ ⁴⁸ | -1.96 | -3.50 | 1 |
| $\text{Li}_{10}\text{GeP}_2\text{S}_{11.7}\text{O}_{0.3}$ ⁴⁹ | -1.99 | -3.06 | 1 |
| $\text{Li}_{10}\text{GeP}_2\text{S}_{11.4}\text{O}_{0.6}$ ⁴⁹ | -2.07 | -3.07 | 1 |
| $\text{Li}_{10}[\text{Si}_{0.3}\text{Sn}_{0.7}]\text{P}_2\text{S}_{12}$ ⁴⁸ | -2.09 | -2.66 | 1 |
| $\text{Li}_{9.42}\text{Si}_{1.02}\text{P}_{2.1}\text{S}_{9.96}\text{O}_{2.04}$ ⁵⁰ | -3.49 | -3.67 | 1 |
| $\text{Li}_{3.35}\text{P}_{0.93}\text{S}_{3.5}\text{O}_{0.5}$ ⁵¹ | -4.04 | -2.67 | 1 |
| $\text{Li}_{3.3}\text{SnS}_{3.3}\text{Cl}_{0.7}$ ⁵² | -4.49 | -3.62 | 0 |
| $\text{Li}_{4.3}\text{AlS}_{3.3}\text{Cl}_{0.7}$ ⁵³ | -5.09 | -7.14 | 0 |
| $\text{Li}_3\text{P}_5\text{O}_{14}$ ⁵⁴ | -6.04 | -7.73 | 0 |
| LiAlP_2O_7 ⁵⁵ | (very low) | -6.32 | 0 |

As there are more training data available than in the validation investigations, the final models should have similar or improved performance to the results observed through cross-validation. The final classification model predicts whether the compounds of the experimental holdout set possess high ($\log_{10}(\sigma) > -4$) or low ionic conductivity with an accuracy of 0.91 and a MCC of 0.83. The final regression model achieves an MAE of 1.34 on the holdout set, with an R^2 score of 0.51. The performance of the final model

against this necessarily small holdout set is consistent with the more robust performance indicators obtained from the previous validation investigations.

Despite the disparity in chemistries between the majority oxide training set and the more varied experimental holdout set, it appears from these metrics and also from consideration at the level of individual materials, that the regressor predicts properties reasonably. Compositions with exceptionally high conductivity are underestimated by the regression model. For nine of the eleven materials, the conductivity has been correctly predicted within two orders of magnitude, which would be expected for materials related to $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$, as this is contained in the training data. However, for the non-oxide materials of the holdout set that are dissimilar to those in the training set, performance is reasonable even when these materials have crystal structures that differ from other materials included in the training set.

$\text{Li}_{3.3}\text{SnS}_{3.3}\text{Cl}_{0.7}$ is the first lithium ion conducting defect stuffed wurtzite based on hexagonal close packed S^{2-} anions⁵². $\text{Li}_3\text{P}_5\text{O}_{14}$ has an ultraphosphate crystal structure defined by extended anionic layers, and is also structurally distinct from materials included in the training set⁵⁴. Given that these are structurally differentiated materials, the ionic conductivities have been reasonably predicted (within 1.69 of the true $\log_{10}(\sigma)$) by a regression model that is based purely on composition. These models can be used as screening tools to motivate the further study of candidate materials and phase fields, and assist in the prioritisation of resource commitment for experimental synthetic work.

Given the intended purpose as a screening tool, and the more favourable metrics demonstrated by the classification model, a reliable classification of high conductivity materials is more helpful than an absolute estimate of the ionic conductivity from the regressor. There are fewer materials with exceptionally high or low conductivity in the database, and as such there will be greater

uncertainty when predicting a specific conductivity for materials in these extrema. Training on classification features gives a more balanced distribution of positive and negative class labels, which gives the model a less skewed dataset for judging its composition-based decision boundary, as reflected in the more favourable performance scores of the classification models. Although there is identified anthropogenic bias present in the dataset, the MCC score under LOCO-CV improves in comparison to each control. This leads us to conclude that these classification models predict with sufficient reliability whether a material has a $\log_{10}(\sigma) > -4$ for these to be further employed to screen candidate ionic conductors (*e.g.*, the material contains Li and is likely to have low electronic conductivity). This does not replace expert chemical knowledge and judgement, instead providing a complementary numerical insight based on evaluation of data at a scale hard for human experts to assimilate.

Conclusion

We present the largest currently available dataset of experimentally reported lithium solid state electrolytes. This dataset includes the composition, structural type, conductivity, and measured temperatures of 789 ACIS measured conductivities, with 403 unique compositions with an associated ionic conductivity near room temperature. Multiple stages of data validation were carried forward by a team of domain experts to ensure that all data are correctly imported from the literature. The creation of a reliable database is a task that is particularly difficult to carry forward with automated tools due to the wide inconsistencies in how data is reported in the field of ionic conductors, necessitating lengthy human validation. Automated scraping would be a viable strategy if all future reports were to prominently state in the abstract a well-defined composition, ionic conductivity in common and clearly stated units (*e.g.* S cm^{-1}), the temperature at which it was measured (*e.g.* 298 K) and the technique used to measure it (*e.g.* ACIS). With

this in mind, we encourage researchers and journal editors to consider reporting core findings in this manner, which will enable materials science researchers to leverage tools from the natural language processing community to gather even larger datasets in the future.

The dataset represents the diversity of chemistry spanned by lithium-containing materials, with a numerical preponderance of oxide-based examples. There are 15 structural families represented at room temperature, including oxides, sulphides, halides, and mixed anion materials. These room temperature compositions are visualised and clustered with the ELM2D package to partition the dataset into nine chemically distinct clusters for leave one cluster cross-validation (LOCO-CV) assessment of the performance of machine learning models.

Supervised statistical (AutoSklearn) and deep learning (CrabNet) models have been applied to this dataset to predict the ionic conductivity of a material from its elemental composition alone. Regression and classification models have been evaluated with standard statistical metrics under different cross-validation regimes to assess their performance at predicting the ionic conductivities of novel materials. The ionic conductivity of a material is the product of many chemical and structural considerations, and also depends on external factors such as temperature. Further, the measured conductivity can also strongly depend on sample preparation, the presence of impurity phases, and crystallite size distribution, which are often discussed collectively under the nebulous term, “sample quality.” This makes ionic transport a difficult property to reliably predict from limited and anthropogenically biased compositional data. Given this challenge, we go beyond standard statistical metrics by designing control studies to more thoroughly investigate the models. We show that CrabNets with transfer learning demonstrate the best performance under both k -folds and LOCO cross-validation.

We present a classification model that is able to estimate whether a material has high or low conductivity which demonstrates improved reliability over random sampling. This is a practical tool to aid experimentalists in their decisions to prioritise candidates for further investigation as lithium ion conductors. Predictions from this model for chemistries dissimilar to those contained in the database are likely to be less reliable than those of closer chemistries, and materials that may have received a low conductivity prediction from these models may still be of interest. This emphasises the importance of reporting newly synthesized materials with distinct chemistry and their measured properties. This should be encouraged even if said property is not seen as being “exceptional” in comparison to heavily investigated and optimised materials families that have seized the attention of many researchers.

Acquiring new data is the only route to improving the performance of supervised models in outlier conductivity regions.

Diversification of the structure and bonding within studied ionic conductors expands the predictive utility of these models because the database on which they are trained is more representative. This experimental synthetic exploration of uncharted chemical (composition and structure) space to generate new examples is thus of foundational importance, regardless of the absolute performance of the arising material. Each qualitatively distinct material in terms of differentiated structure and bonding assists our understanding of where high performing materials may be located in chemical space. This distinguishes the generation of materials closely related to existing examples -- which is valuable for optimisation -- from studies that explore distinct parts of the relevant chemical space. The model performance here reinforces the importance of exploratory discovery synthesis coupled with definition of structure-property-composition relationships for lithium ion transport.

References

1. Goodenough, J. B. Rechargeable batteries: challenges old and new. *J. Solid State Electrochem.* **16**, 2019–2029 (2012).
2. Knauth, P. Inorganic solid Li ion conductors: An overview. *Solid State Ion.* **180**, 911–916 (2009).
3. Janek, J. & Zeier, W. G. A solid future for battery development. *Nat. Energy* **1**, 1–4 (2016).
4. Wang, Y. *et al.* Design principles for solid-state lithium superionic conductors. *Nat. Mater.* **14**, 1026–1031 (2015).
5. Bachman, J. C. *et al.* Inorganic Solid-State Electrolytes for Lithium Batteries: Mechanisms and Properties Governing Ion Conduction. *Chem. Rev.* **116**, 140–162 (2016).
6. Lombardo, T. *et al.* Artificial Intelligence Applied to Battery Research: Hype or Reality?
<https://doi.org/10.1021/acs.chemrev.1c00108>. *Chem. Rev.* (2021) Articles ASAP.
7. Sendek, A. D., Cheon, G., Pasta, M. & Reed, E. J. Quantifying the search for solid Li-ion electrolyte materials by anion: a data-driven perspective. *J. Phys. Chem.* **124**, 8067–8079 (2020).
8. Zhang, Y. *et al.* Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **10**, 5260 (2019).
9. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
10. Haghighatlari, M., Shih, C.-Y. & Hachmann, J. Thinking Globally, Acting Locally: On the Issue of Training Set Imbalance and the Case for Local Machine Learning Models in Chemistry. Preprint at <https://chemrxiv.org/engage/chemrxiv/article-details/60c745c4337d6cef32e2704f>. (2019).
11. De Breuck, P.-P., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys. Condens. Matter* **33**, 404002 (2021).
12. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *Npj Comput. Mater.* **6**, 1–10 (2020).
13. Irvine, J. T. S., Sinclair, D. C. & West, A. R. Electroceramics: Characterization by Impedance Spectroscopy. *Adv. Mater.* **2**, 132–138 (1990).

14. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
15. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5**, 180111 (2018).
16. Sendek, A. D. *et al.* Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
17. Fergus, J. W. Ceramic and polymeric solid electrolytes for lithium-ion batteries. *J. Power Sources* **195**, 4554–4569 (2010).
18. Rossbach, A., Tietz, F. & Grieshammer, S. Structural and transport properties of lithium-conducting NASICON materials. *J. Power Sources* **391**, 1–9 (2018).
19. Stramare, S., Thangadurai, V. & Weppner, W. Lithium Lanthanum Titanates: A Review. *Chem. Mater.* **15**, 3974–3990 (2003).
20. Zhang, Z. *et al.* New horizons for inorganic solid state ion conductors. *Energy Environ. Sci.* **11**, 1945–1976 (2018).
21. Chen, C. & Du, J. Lithium Ion Diffusion Mechanism in Lithium Lanthanum Titanate Solid-State Electrolytes from Atomistic Simulations. *J. Am. Ceram. Soc.* **98**, 534–542 (2015).
22. Xiang, Y.-X. *et al.* Toward understanding of ion dynamics in highly conductive lithium ion conductors: Some perspectives by solid state NMR techniques. *Solid State Ion.* **318**, 19–26 (2018).
23. Nolan, A. M., Zhu, Y., He, X., Bai, Q. & Mo, Y. Computation-Accelerated Design of Materials and Interfaces for All-Solid-State Lithium-Ion Batteries. *Joule* **2**, 2016–2046 (2018).
24. Manawan, M., Kartini, E. & Avdeev, M. Visualizing lithium ions in the crystal structure of Li_3PO_4 by in situ neutron diffraction. *J. Appl. Crystallogr.* **54**, 1409–1415 (2021).
25. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners. *OpenAI Blog* https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
26. Wolf, T. *et al.* Transformers: State-of-the-Art Natural Language Processing. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020).
27. Hargreaves, C. J., Dyer, M. S., Gaultois, M. W., Kurlin, V. A. & Rosseinsky, M. J. The Earth Mover’s Distance as a Metric for

- the Space of Inorganic Compositions. *Chem. Mater.* **32**, 10610–10620 (2020).
28. Levin, I. NIST Inorganic Crystal Structure Database (ICSD). <https://doi.org/10.18434/M32147>. (2020).
29. Krzanowski, W. in *Principles of Multivariate Analysis Ch. 2* (Oxford Univ. Press, Oxford, 2000).
30. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).
31. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at <https://arxiv.org/pdf/1802.03426.pdf>. (2018).
32. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).
33. Murdock, R. J., Kauwe, S. K., Wang, A. Y. T. & Sparks, T. D. Is Domain Knowledge Necessary for Machine Learning Materials Properties? *Integrating Mater. Manuf. Innov.* **9**, 221–227 (2020).
34. Wang, A. Y.-T. *et al.* Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **32**, 4954–4965 (2020).
35. Ho, T. K. Random decision forests. in *Proceedings of 3rd International Conference on Document Analysis and Recognition* vol. 1 278–282 vol.1 (1995).
36. He, X., Zhao, K. & Chu, X. AutoML: A Survey of the State-of-the-Art. *Knowl.-Based Syst.* **212**, 106622 (2021).
37. Feurer, M. *et al.* Efficient and Robust Automated Machine Learning. in *Proceedings of the 28th International Conference on Neural Information Processing Systems* vol. 2 2755–2763 (2015).
38. Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *Npj Comput. Mater.* **7**, 1–10 (2021).
39. Vaswani, A. *et al.* Attention is All you Need. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
40. Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).
41. Kong, S., Guevarra, D., Gomes, C. P. & Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **8**, 021409 (2021).

42. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Comput. Mater.* **1**, 1–15 (2015).
43. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. *MIT Press* (2016).
44. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta BBA - Protein Struct.* **405**, 442–451 (1975).
45. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
46. Stanev, V. *et al.* Machine learning modeling of superconducting critical temperature. *Npj Comput. Mater.* **4**, 1–14 (2018).
47. Jiang, Y., Hu, Z., Ling, M. & Zhu, X. A comparative study of $\text{Li}_{10.35}\text{Ge}_{1.35}\text{P}_{1.65}\text{S}_{12}$ and $\text{Li}_{10.5}\text{Ge}_{1.5}\text{P}_{1.5}\text{S}_{12}$ superionic conductors. *Funct. Mater. Lett.* **13**, 2050031 (2020).
48. Sun, Y., Suzuki, K., Hori, S., Hirayama, M. & Kanno, R. Superionic Conductors: $\text{Li}_{10+\delta}[\text{Sn}_y\text{Si}_{1-y}]_{1+\delta}\text{P}_{2-\delta}\text{S}_{12}$ with a $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ -type Structure in the Li_3PS_4 – Li_4SnS_4 – Li_4SiS_4 Quasi-ternary System. *Chem. Mater.* **29**, 5858–5864 (2017).
49. Sun, Y. *et al.* Oxygen substitution effects in $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ solid electrolyte. *J. Power Sources* **324**, 798–803 (2016).
50. Hori, S., Suzuki, K., Hirayama, M., Kato, Y. & Kanno, R. Lithium Superionic Conductor $\text{Li}_{9.42}\text{Si}_{1.02}\text{P}_{2.1}\text{S}_{9.96}\text{O}_{2.04}$ with $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ -Type Structure in the Li_2S – P_2S_5 – SiO_2 Pseudoternary System: Synthesis, Electrochemical Properties, and Structure–Composition Relationships. *Front. Energy Res.* **4**, 38 (2016).
51. Suzuki, K. *et al.* Synthesis, structure, and electrochemical properties of crystalline Li–P–S–O solid electrolytes: Novel lithium-conducting oxysulfides of $\text{Li}_{10}\text{GeP}_2\text{S}_{12}$ family. *Solid State Ion.* **288**, 229–234 (2016).
52. Vasylenko, A. *et al.* Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nat. Commun.* **12**, 5561 (2021).
53. Gamon, J. *et al.* $\text{Li}_{4.3}\text{AlS}_{3.3}\text{Cl}_{0.7}$: A Sulfide–Chloride Lithium Ion Conductor with Highly Disordered Structure and Increased Conductivity. *Chem. Mater.* **33**, 8733–8744 (2021).
54. Han, G. *et al.* Extended Condensed Ultraphosphate Frameworks with Monovalent Ions Combine Lithium Mobility with High Computed Electrochemical Stability. *J. Am. Chem. Soc.* **143**, 18216–18232 (2021).

55. Shoko, E. *et al.* Polymorph of LiAlP_2O_7 : Combined Computational, Synthetic, Crystallographic, and Ionic Conductivity Study. *Inorg. Chem.* **60**, 14083–14095 (2021).

Organising Materials Datasets for Similarity Queries with Metric Indexing Data Structures

The concluding section of this work on the EIMTree is under preparation for publication as part of a wider manuscript for the suite of tools hosted on the Liverpool Materials Discovery Server (LMDS), co-written by Samantha Durdy, myself, Michael W. Gaultois, and Matthew S. Dyer. The LMDS project was conceived and executed by myself and Matthew S. Dyer, with all installation and technical assistance provided by Mark Dennison and Ben Wagg of the university of Liverpool servers and storage team. It was funded as part of the SOLBAT grant FIRG007, with support from Matthew J. Rosseinsky.

Introduction

Searching for things is the preliminary task in many daily processes. As we awake we search for the alarm, in the shops we search for our favourite foods, and in our work we search for similar materials to those under investigation. In the general problem, we have a given set of objects that we are trying to search against, such as things in our immediate reach, items in the shop, and databases of materials. We also have a query. These may be well formed, such as the position of the alarm clock or less well formed, being the list of things you think you want to eat. These queries may be entirely subjective, such as a more similar material. More similar compositionally? More similar structurally? More similar in a specific property? Just *more* similar?

Comparing the outputs of generative models against reported databases of compositions to find their most similar match is an excellent application of the EIMD, where the models output is the query. Unfortunately it can be very computationally costly to search many databases in a high throughput manner. Efficiently

navigating these spaces enables researchers to compare potential formulations against a greater body of past research, which may prevent the duplication of efforts to synthesize a novel material.

The use of imprecise search tools may not highlight areas of interest, or incorrectly identify objects as being similar when others may be better matches. A common difficulty is that if objects are unsorted and scrambled amongst one another, then each must be investigated in turn until we find the desired, or closest, object that we are searching for. This brute force method of investigation is very time consuming, especially if the method of comparison is computationally costly. A better approach is to structure objects logically so they may be accessed through an indexing scheme.

To our knowledge, there are no prior indexing schemes reported for materials datasets, and many composition based interfaces rely on entering the composite elements individually. This means that searches are slow to enter for the user, relying on imprecise database queries. A simple approach could store each database entry with its chemical formula as a composition vector. Chemical queries could be searched for in a brute force manner, by comparing each of the stored vectors with the query vector using the compositional Euclidean distance (CED) to score similarity. This has the same drawbacks discussed in Chapter 2, but unfortunately, an EIMD based brute force search becomes computationally costly leading to unacceptably long search times.

To index objects for searching we must use a data structure. Some data structures can be very simple and efficient in their use cases, such as arrays of objects that are accessed by an index. For example, an elevator does not need to spend any time searching for the correct floor, as each button press maps to the correct index to load the instruction. This is not an informative search, nor should it be, as we do not wish the elevator to visit the five closest floors as it searches for the desired floor. For more complex problems that involve indexing sets of objects with numeric representations with irregular intervals between members, or when performing non-

exact similarity queries, we need more advanced data structures to store and execute these efficiently. These are typically introduced in an undergraduate computer science course, but an introduction to this topic with some common examples is provided in Appendix 1.

Unfortunately, the most commonly used search trees are unsuitable for the purpose of indexing compositions *w.r.t.* the EIMD, as these are typically reliant on the Euclidean distance, *i.e.* the CED. By using the CED these data structures enforce the use of a coarse metric between vectors, which leads to greatly reduced distinguishing power. This can mean that highly dissimilar objects will be found in local subtrees of the datastructure which means time is wasted computing these distances when they could be ignored. Further, this data structure is reliant on each indexing object being a high dimensional fixed length vector. The representations generated by Percifter in the next chapter are of variable length, which means that we cannot use many of these trees to index Percifter objects. As materials do not come with an associated frame of reference such as Euclidean coordinates, we cannot order these spatially. Thus, we rely solely on the information given to us from the metric.

Thankfully, the field of data structures called metric indexing trees focuses on this specific problem. When organising the space, objects that are in local proximity to one another in the metric space should be located in subtrees in close proximity to one another. This allows us to reduce the number of distance computations which must be made when performing a search. This will be of particular significance for computationally expensive metrics, such as the EMD, where brute force searching may be time prohibitive.

Here we will demonstrate how the previously investigated metrics may be applied to structure materials datasets, allowing us to navigate these spaces quickly and efficiently. There may not be a consistent ordering that can be applied to an arbitrary metric space.

However, the machinery that metric spaces offer, in particular the triangle inequality, can be applied to consistently index objects into hierarchical search trees of candidate objects. There are many metric indexing trees^{3,4}, and in this chapter we investigate the use of the MTree⁵ and the List of Clusters⁶ (LC) approaches to indexing spaces of materials *w.r.t.* the EIMD distances⁷.

Metric Spaces

Partitioning spaces to execute similarity searches will of course require some form of similarity score. Objects of the space must have an associated metric, a low value indicating similarity between objects, and larger values dissimilarity. Metrics are introduced in Chapter 1 and discussed in later chapters with applications of the EMD, but the importance of these rules to the indexing scheme permits us to reiterate. A metric, d , between two objects, O , must satisfy the metric axioms⁸:

- 1) $d(O_x, O_y) = 0, O_x = O_y$
- 2) $d(O_x, O_y) \geq 0, O_x \neq O_y$
- 3) $d(O_x, O_y) = d(O_y, O_x)$
- 4) $d(O_x, O_z) \leq d(O_x, O_y) + d(O_y, O_z)$

The first of these rules (Figure 1(a)) states that if the distance between two objects is 0, then these must be the same object under the given metric. The second (Figure 1(b)) is the property of non-negativity, for if two identical objects have a distance of 0, then it cannot be possible for objects to be *more* identical and thus return a negative score. The third of these rules (Figure 1(c)) is required for consistency, stating that a metric should be a symmetric function, *i.e.*, the ordering of the input objects should not impact the output value of the metric function.

The fourth, the triangle inequality (Figure 1(d)), states that the distance between any two objects must be less than or equal to the length of the path between them that visits an intermediate object, which is to say, the shortest path between two points is always an uninterrupted line. The triangle inequality is the core tool that will

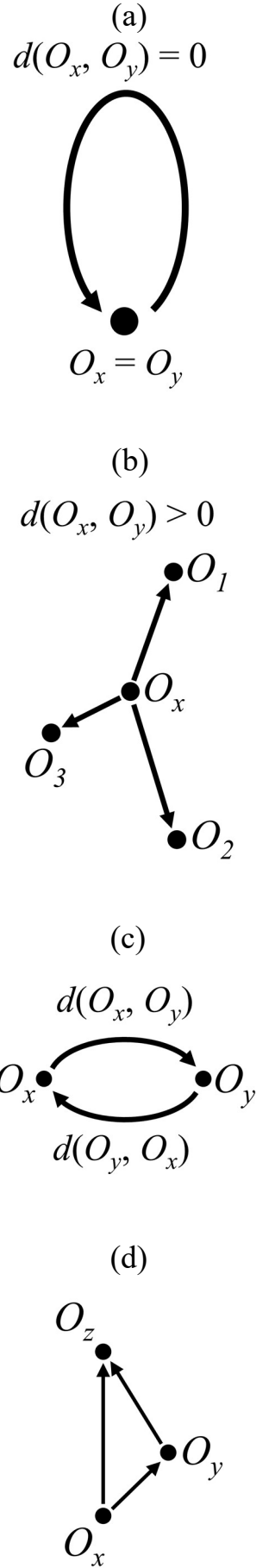


Figure 1: Visual representations of each of the metric axioms in two-dimensional space.

be used to prune entries from our search when performing similarity queries on the MTree and LC.

Metric Trees

Objects may not have easily identifiable features, but if there is a metric between objects, we can use this information to index objects with respect to their distance to other objects. There are two core approaches to construct metric indexing trees, generalized hyperplane partitioning and ball partitioning methods. Each of the metric tree algorithms apply these principles differently, and sometimes in combination, to partition spaces. Below we outline a general overview of these methods applied to 2D Euclidean co-ordinates.

In generalized hyperplane partitioning schemes (Figure 2), when each point is indexed, it is assigned to a child, dependent on which of the two children, called the pivots, the point falls closest to. This has the effect of inserting a plane following the medial line that is orthogonal to the two pivots. This is called a generalised hyperplane, which will partition the space in two. During queries, the points will be routed depending on which of the two pivots they fall closest to at each level of the tree. This tends to create more symmetric and balanced indexing trees, which may mean that more distance calculations will need to be made for irregularly distributed spaces.

Rather than using deep learning, the popular music streaming platform spotify uses a generalized hyperplane indexing tree called Approximate Nearest Neighbours Oh Yeah (ANNOY)⁹ for its music recommendation engine. User embeddings are created through a matrix factorisation of the network of artists that they listen to, with the Euclidean distance between embeddings used as the metric. Instead of choosing the hyperplane as the plane that falls midway between two points, ANNOY uses randomly projected hyperplanes at each split. Each user is paired to their most similar vectors in the database, and their new musical suggestions taken from the artists that similar users listen to.

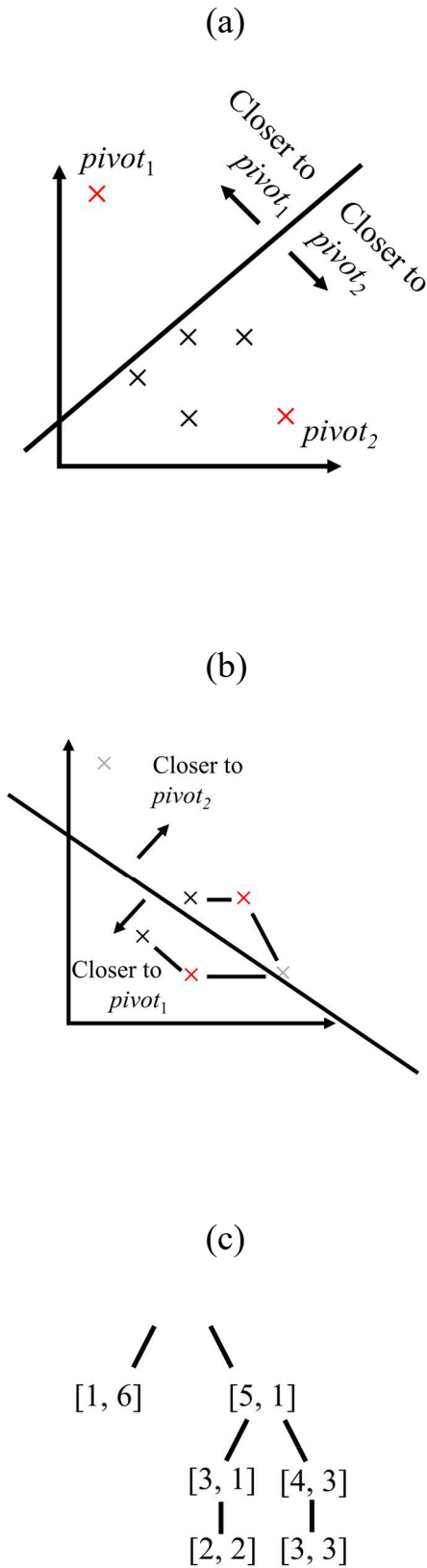


Figure 2: Six 2-dimensional points are indexed via the generalised hyperplane method. Hyperplanes are overlaid in (a) and (b) with the resultant metric tree given in (c).

Ball partitioning methods, sometime called Ball Trees (Figure 3), also rely on indexing the dataset based on the distances to particular objects chosen from the indexing set, which are now referred to as vantage points (VP). Every object of the indexing dataset is stored in a node of the tree, with vantage points contained in the internal nodes. Vantage points may be chosen from the dataset at random, assigned based on regions of local density, or updated dynamically as new objects are added to the tree. As well as the nodes routing object, vantage points store an additional attribute at each node, which is the covering radius. This is equal to the distance between the vantage point and the furthest object contained in its subtree, which will be updated during construction as the subtree grows. The covering radius is the property that enables us to prune regions of the tree during similarity searches, through the triangle inequality (Figure 4). Let us illustrate this with a simple analogy.

You have decided that if there is one within an acceptable proximity radius, $r(Q)$, of 1,000 km to your house, you would like to visit a space ship. You have acquired a metric indexing tree that has all the space vehicles of the solar system indexed using a ball partitioning scheme with respect to the Euclidean distance. The root node contains pointers to two entries, a rocket at the Kennedy space centre in Florida (O_{r1}), and the perseverance rover on mars (O_{r2}), shown in Figure 5.

You query the tree using radio waves sent from your house, Q , to find the distance to the routing objects. In doing so, the distance between the query object Q and each of the routing objects, O_{r1} , and O_{r2} , is measured to see whether Q falls inside the covering radius of either of the subtrees. The covering radius of the perseverance rover, $r(O_{r2})$, is 10,000 km, and the distance $d(Q, O_{r2})$ is 25,000,000 km. Whilst there are fifteen entries indexed under O_{r2} , one for each of the mars missions, the objects of the subtree have to be at least as close to O_{r2} than the covering

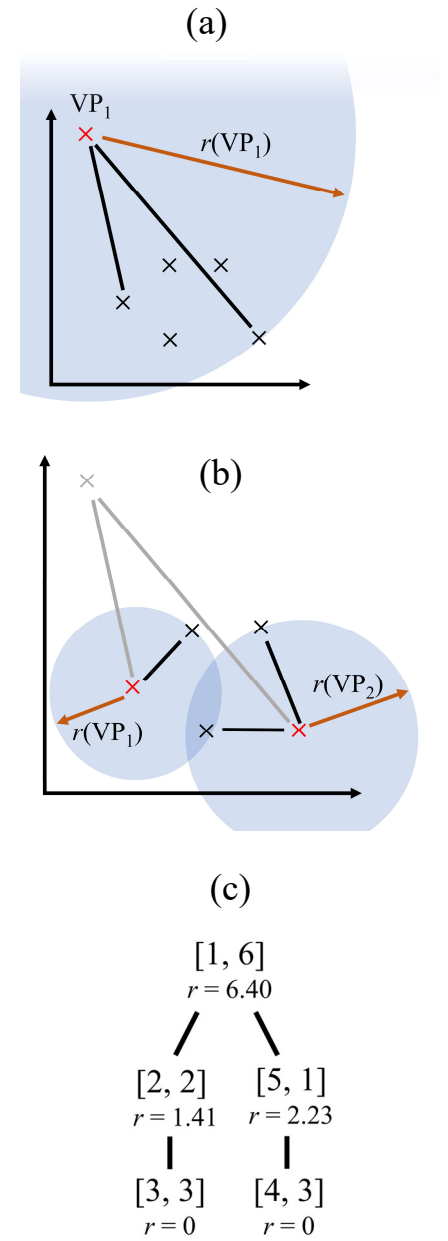


Figure 3: The points of Figure 2 are indexed via a ball-partitioning method. Covering disks are overlaid in (a) and (b) with the resultant metric tree given in (c).

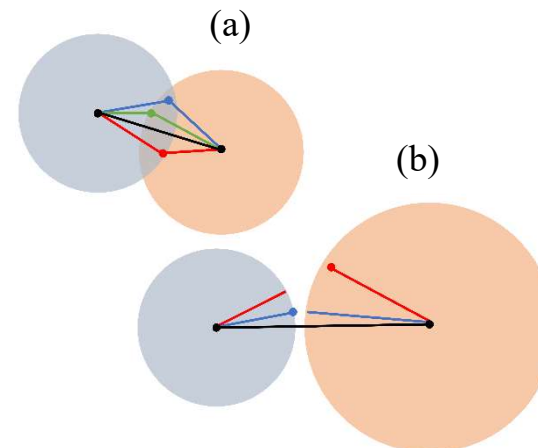


Figure 4: Each of the query points in (a) fall under both vantage points covering disks so no subtree may be pruned. In (b) each of the opposing subtrees may be removed from the search

radius of 10,000 km. The closest that any of the Martian objects could physically be to Q is thus 24,990,000 km.

As the distance to the Kennedy centre, $d(Q, O_{r1})$, is 6,600 km, and this falls inside the covering radius $r(O_{r1})$ of 18,000 km, one of the entries in the subtree may be close enough, as $d(Q, O_{r1}) < r(O_{r1})$. The children of $T(O_{r1})$ must be closer to Q than the children of $T(O_{r2})$, so we can prune the tree $T(O_{r2})$ from our search entirely and descend to the subtree $T(O_{r1})$. As time-of-flight radio waves were used to measure distance, at no point were the locations of any of these objects required. Just the information provided through the metric.

Here it is clear to see that we could have a close match with one of the descendants of O_{r1} , as the disk of acceptable proximity completely intersects the region of indexed objects, but we cannot tell for sure until we descend the tree and check. In general, the geometric principles of the triangle inequality may be used to prune subtrees if we have an acceptable proximity radius, a distance to a routing object, and the routing object covering radius.

With O_{r1} , any point could be chosen at random within the acceptable proximity radius of Q . This point would be guaranteed to be acceptably close to Q , intersect with the indexing region of O_{r1} thus potentially being contained in O_{r1} , and would form a valid triangle with O_{r1} and Q . With O_{r2} however, it would not be possible to construct a triangle with edge lengths 1,000, 25,000,000, and 10,000. This tells us there could not be an object of interest in that region of the metric space which satisfies the search criteria without breaking the triangle inequality. There are other cases, such as when a routing object has a distance of 0, in which case we have an exact match. If the disk of proximity partially intersects with a routing object covering disk, then the subtree cannot contain an exact match, but could contain objects of interest.

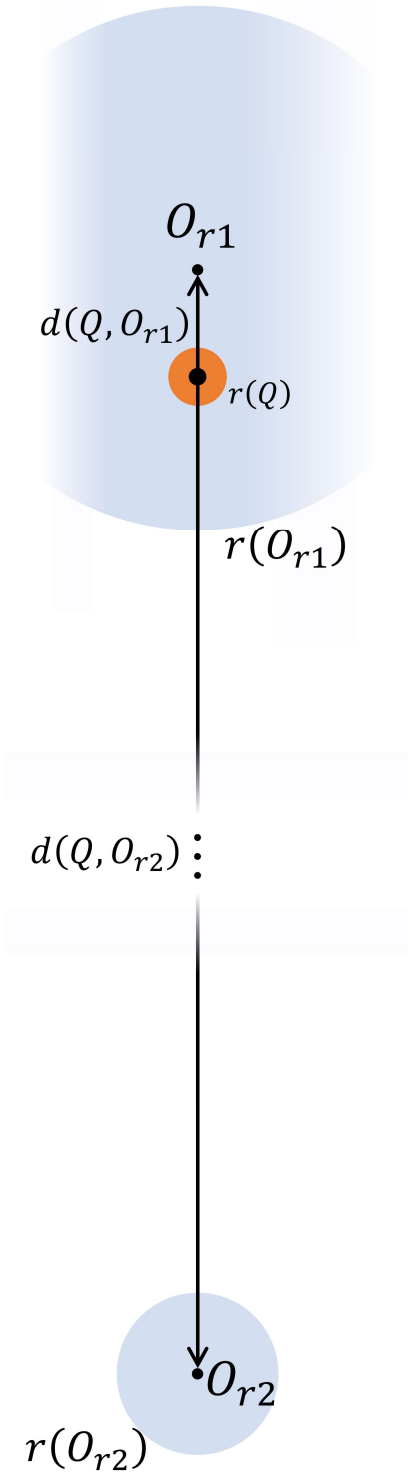


Figure 5: The distance $d(Q, O_{r1})$ from our house, Q , and the Kennedy centre, O_{r1} , falls inside the covering radius of the Kennedy centre so we should investigate this subtree. The distance $d(Q, O_{r2})$ to the perseverance rover O_{r2} given by the dashed line is much greater than the covering radius of O_{r2} so we can prune this subtree

Ball partitioning schemes are beneficial as they often generate asymmetric and imbalanced trees which are good at partitioning datasets of irregular density and are equipped with robust pruning schemes. However, the covering radii of vantage points can overlap, which means that more subtrees must be processed than necessary when executing queries. This effect will be even more pronounced on evenly distributed spaces with consistent density, as there will be many vantage points with overlapping covering radii.

Priority Queues

An additional data structure which will be used in this chapter is the priority queue ¹. This is a simple construction, which stores a list of objects, O , each with an associated key, k , in an array:

$$[(k_1, O_1), (k_2, O_2), \dots (k_n, O_n)]$$

Each of the entries are ordered by the value of k , such that $k_1 < k_2 \dots < k_n$. When a new object is added to the queue, the appropriate position for the new key is found and the object inserted into this position. As the queue is already sorted, this operation can be performed in $O(\log(n))$ time by applying a binary search tree operation. When working with distances to queries, this means that the item in the first position of the queue will always contain the closest object that has been considered, and the last item in the queue will be the object furthest away.

In this chapter, three priority queues are used whilst indexing and search processes are executing. The first is a nearest neighbours' queue, NN, which stores the closest matches to a query and is returned once the algorithm has terminated (Figure 6). The second is used by the MTree during k -NN searches and is a tree queue, TQ, which stores pointers to each of the subtrees in the MTree that are yet to be examined. In TQ, the distance between the query and the furthest object in the subtree is used as the key. The third is used by the list of clusters (LC) to store the points stored under each vantage point, sorted by their distance to the vantage point.

MTree.knn(NaCl, 5):

| | |
|---------|-----------------------------------|
| (0, | NaCl), |
| (0.05, | $K_{0.05}Na_{0.45}Cl_{0.5}$), |
| (0.051, | $Na_{0.5}Br_{0.051}Cl_{0.449}$), |
| (0.088, | $K_{0.088}Na_{0.412}Cl_{0.5}$), |
| (0.1, | $K_{0.1}Na_{0.4}Cl_{0.5}$) |

Figure 6: The final five entries of the NN priority queue when executing an EIMD similarity search for NaCl on either the MTree and List of Clusters. These are given as a list of ordered pairs (distance to query, composition) and sorted by distance to the query.

The MTree

An indexed MTree stores each of the entries in the leaves of the tree, with duplicates of each object stored further up the tree in internal nodes to route queries to the nearest leaf node. The tree first has a root node, containing the feature information of an associated routing objects, O_r . Each node contains pointers to child routing objects, each of which will also have associated object information, and further sub-trees if they are non-leaf nodes. The subtree associated with a node object is denoted $T(O_r)$. We need not limit ourselves to binary trees, which is to say trees where each node has two children, and many pointers can be stored at each vantage point. This extension is called the n -ary tree, which can take up to n pointers at each node. If a node exceeds this limit, we say that it overflows, and we must execute a costly promotion scheme to rebalance the tree.

In addition to storing object information, each internal routing object operates as a vantage point, containing the radius of the disk, $r(O_r)$ which covers all entries contained in $T(O_r)$, *e.g.* the distance from the routing object to the furthest object contained in its children and sub-children. Each non-root object also stores the distance between itself and its parents routing object, $d(O_r, O_p)$ to save repeated calculation.

We assume that each of the objects to be indexed are entered randomly, with no guarantee of proximity. The tree must therefore be rebalanced during construction to ensure that the vantage point objects span representative regions of the space, to efficiently route query objects to the correct matches. Objects are added in a bottom-up fashion, where each new object is compared with the root, and recursively descended to a suitable leaf node until the distance between the new object falls outside the covering radius of a parent routing object, whereupon a new leaf is created.

When the new object falls inside the covering radius of multiple routing objects then it is routed to the closest object. When the new

object reaches a leaf node or a node where the distance $d(O_n, O_r)$ between the new object, O_n , and each routing object O_r is greater than the covering radius $r(O_r)$, then the new object is added to the closest of the possible routing objects. If this causes a node to exceed its capacity, a node split is triggered, and a routing object is promoted to a higher level up the tree. This may trigger another internal node to split, and the process will continue until a new root node has been promoted, or a routing object is added to an internal node below capacity.

An ideal MTree would minimise the total sum of the covering radii, with little overlap between covering radii, as this allows us to prune a greater number of subtrees during searches. The volume of the tree can be computed by summing together all covering radii, where a minimal volume ensures the tree is not unnecessarily deep, with well clustered trees. Different promotion policies are provided in the original publication, with the minimum sum of radii (m_RAD) policy optimally reducing the overall volume of the tree.

m_RAD is a computationally intensive procedure, which constructs a complete distance matrix of all possible combinations of objects in the tree to recursively pick the two objects which minimise the total volume. Due to time constraints when indexing MTrees with large datasets, the Maximum Lower Bound on Distance (M_LB_DIST) promotion scheme is instead used in this work. This involves forming a distance matrix between the objects in the node under consideration and selecting the two objects which are furthest away from one another to promote up a level. Whilst this leads to less efficient search trees, empirical search times of the implementation are found to be acceptable.

Once the two objects to promote have been decided, each of the sub-trees must be assigned to a new parent routing object. A perfectly balanced tree can be formed by sequentially selecting the closest routing objects to each of the parent nodes in turn. This will ensure equal numbers of entries are contained in the sub-trees at each split, which minimises the total number of routing nodes

required to construct the tree. However, this leads to less clustered trees that do not exploit the local densities of the objects of the metric space, thus leading to a greater overall tree volume. Instead, the generalised hyperplane-partitioning scheme is applied here, which simply assigns each of the objects to their most similar parent, leading to less balanced trees.

When all objects have been indexed, we may determine which objects fall within the covering radius of a query object, O_q , or determine the k -nearest neighbours (k -NN) of a query object. Given the intended purpose as a direct matching tool, we will detail the k -NN search, and refer to the earlier outline for the theory of the range search. Both have been implemented in the EIMTree codebase.

k -NN Search of an indexed MTree

An indexed MTree stores the objects of a space in the leaves of a tree data structure. Each internal node of the MTree stores multiple routing entries, the number of internal routing entries to store at each node being a hyper parameter that is set before indexing. The binary variant with two entries per node is discussed before presenting the performance in search times of n -ary trees in the results section. Each entry stores the object information and pointers to routing objects that are its children; if it is a leaf node it stores only the object. Further to this, each internal node stores the value of a covering radius, which is calculated as the largest distance between a nodes routing object and every object contained in its subtree, as well as the distance to its parent if it is not the root of the tree.

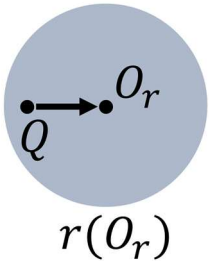


Figure 7: An object O_r has an associated covering radius $r(O_r)$. As $d(Q, O_r) < r(O_r)$, d_{min} is set to zero for the subtree $T(O_r)$, as this subtree may contain an exact match, and this is added to the front of the tree queue (TQ).

Let us formalise this process. We will use two priority queues, NN, which at termination will store the final results of the nearest neighbour search, and TQ. TQ stores pointers to the active subtrees associated with internal routing objects, $T(O_r)$, to continue the search for nearest neighbours. These pointers are ordered by a distance value, d_{min} , which is the closest *potential* distance to an object stored in the subtree. If an object falls inside the covering

radius, $d(Q, O_r) - r(O_r) \leq 0$, this is given the value of 0 and stored at the front of the TQ, as it is possible that the subtree contains a perfect match (Figure 7). Otherwise d_{min} is given the value of $d(Q, O_r) - r(O_r)$, as it is possible that none of the subtrees contain good matches and we simply want the closest values (Figure 8).

$$d_{min}(T(O_r)) = \max \{0, d(Q, O_r) - r(O_r)\}$$

Each entry of TQ will be addressed sequentially, and then removed from TQ, with additional subtrees added to TQ as the MTree is descended. Simply iterating this process would lead to all objects of the tree being queried and sorted in a brute force fashion, but we can use the information from the NN array to prune subtrees from TQ during execution.

NN is a priority queue, containing tuples, $(d(Q, O_n), O_n)$, where O_n is the n -th nearest neighbour to Q that has currently been found. As each object in the nodes of the MTree are queried, if $d(Q, O_i)$ is smaller than $d(Q, O_k)$ then an ordered insertion is performed to enter this into the correct position of NN. The first k objects in NN will be returned by the k -nn procedure. The distance associated with the k 'th nearest object, $NN[k] = (d(Q, O_k), O_k)$ gives an upper bound, $d_k = d(Q, O_k)$, which we may use to define the first of two pruning rules during the search.

If $d_{min}(T(O_r)) > d_k$, the closest possible object in $T(O_r)$ must be further away than the currently identified k -th closest neighbour, and thus the entire subtree may be removed from TQ (Figure 9). This procedure is applied repeatedly each time the value of d_k comes down, without having to compute the distance between Q , and objects in each subtree.

The second pruning rule uses one of the stored properties of each node, as each will contain the distance between its routing object and the parent. This information is used in addition to the covering radius, to give a lower bound on the distance to the closest potential match in a subtree. If this is greater than d_k , then $T(O_r)$

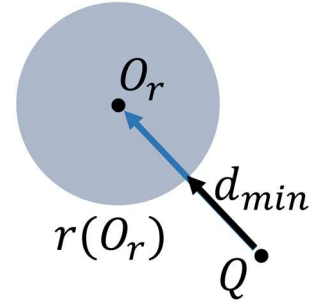


Figure 8: Here $d(Q, O_r) > r(O_r)$ so we calculate the minimum distance an indexed point could theoretically be in $T(O_r)$ and add this to TQ with this value.

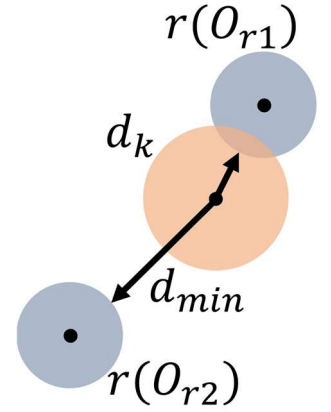


Figure 9: Taking the k 'th distance in NN as d_k , we may prune any subtrees from TQ which have $d_{min}(T(O_r)) > d_k$. In this example $T(O_{r2})$ may be pruned from the search.

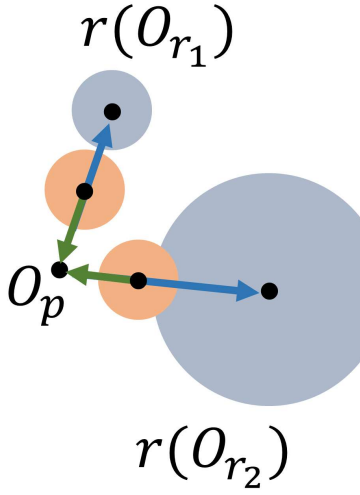


Figure 10: We need only compute the distance between the query and each routing object if the inequality containing each of the precomputed distances holds:
 $|d(Q, O_p) - d(O_r, O_p)| \leq d_k + r(O_r)$
 Here, we would not need to investigate O_{r1} but would add O_{r2} to TQ.

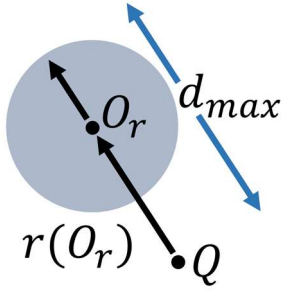


Figure 11: A value of d_{max} may be computed as $d(Q, O_r) + r(O_r)$ and inserted into NN, which may have the effect of lowering d_k .

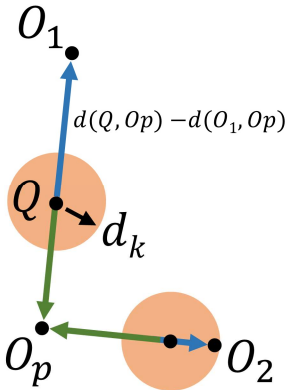


Figure 12: If the difference between the distance between the query and the parent, $d(Q, O_p)$, and the distance between the leaf object and the parent, $d(O_i, O_p)$, is less than d_k then the distance between the query and the leaf is calculated.

cannot contain any objects which are closer to Q than the currently indexed object at $NN[k]$. This means we can avoid computing $d(O_r, Q)$, and the subtree need not be added to TQ for further investigation. Accordingly, a tree $T(O_r)$ will be considered for addition to TQ if the following statement holds (Figure 10):

$$|d(Q, O_p) - d(O_r, O_p)| \leq d_k + r(O_r)$$

Once it is established that the subtree could feasibly contain an object which is a closer match than those in our NN array, the distance $d(Q, O_r)$ is computed, from which $d_{min}(T(O_r))$ is derived. If $d_{min} \leq d_k$, then $(d_{min}, T(O_r))$ is pushed to TQ.

The distance from Q to the furthest possible objects contained in $T(O_r)$, d_{max} , can also be obtained from the given information. The furthest object can be at most the distance to the routing object $d(Q, O_r)$ plus the covering radius (Figure 11).

$$d_{max}(T(O_r)) = d(Q, O_r) + r(O_r)$$

Whilst we do not currently know the identity or exact distance to the furthest object, if $d_{max} \leq d_k$, a placeholder of $(-, d_{max}(T(O_r)))$ may be inserted at the appropriate place of NN. This could have the effect of lowering the value of d_k , as the old value of $NN[k]$ may now be the $k+1$ nearest neighbour. The updated value of d_k allows us to prune further subtrees from TQ, if the associated $d_{min}(T(O_r)) > d_k$.

When a leaf node is encountered, the triangle inequality can be applied to save computing the distance between the query and the leaf object. This is done with the precomputed stored distances of each of the leaf entries, O_i to their parent routing object, O_p , and the distance from Q to O_p . The leaf entries to be considered for entry are those where:

$$|d(Q, O_p) - d(O_i, O_p)| \leq d_k$$

If this holds, the distance $d(Q, O_i)$ is calculated, where once again should $d(Q, O_i) < d_k$, then $(O_i, d(Q, O_i))$ will be inserted into NN,

and TQ pruned with the updated value of d_k . This process is continued until there are no subtrees in TQ with $d_{min} < d_k$, at which point the algorithm terminates and we may return the k closest objects from NN.

The List of Clusters

The MTree was initially chosen due to the generally favourable traits of the data structure that had been reported by the publication. However, recent reviews have shown this gives reduced search performance when compared to other techniques. In particular, the List of Clusters (LC) ⁶ data structure was demonstrated to outperform other methods in retrieval times when benchmarked against high dimensional datasets. After all this discussion of hierarchical partitioning trees, it could appear that this technique takes a deceptively simple approach which would be expected to return inefficient clusters of the space. In practice we have found that not only is this recently implemented method simpler to explain, it is shown to outperform more involved techniques when applied to materials datasets. Let us describe the construction process first.

We choose n/m pivot points, O_p , from the dataset at random, where n is the total number of points and m is the average number of points partitioned to each pivot after construction. In the original list of clusters publication it was recommended to use a value of 16 for m . For our dataset of 1,065,165 compositions this leads to 66,573 pivots. The distance between each of the remaining objects and every pivot is computed. The object is then assigned to the closest pivot, keeping a record of the object and its distance to the parent pivot object. This has the effect of creating a Voronoi partitioning of the space around each of the pivots, or cluster centres, with each point being assigned to the Voronoi cell that it intersects. Each pivot stores the children in a priority queue, sorted by distance to the pivot, taking the greatest distance as a covering radius (Figure 13), allowing for pruning during searches.

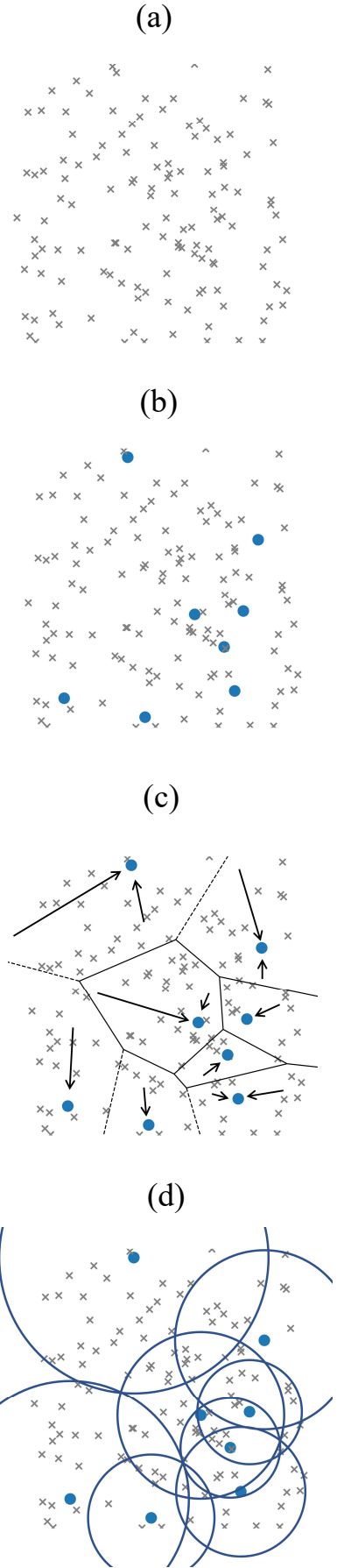


Figure 13: 100 points are shown on the plane (a) with 8 of these chosen as pivots (b). These form Voronoi partitions (c) which will then have associated covering radii attached when considered as vantage points (d).

During searches, the method of operation changes, and each of these pivoting cluster centres are instead treated as vantage points. Once the assignment of each point in the dataset to its closest parent is complete, each pivot takes the distance to its furthest child and stores this as the covering radius, $r(O_p)$. With these routing objects operating as vantage points, pruning the search space similar to the MTree may be carried forward during searches.

A query must first compute the distance to all n/m vantage points. Once this is complete, the vantage points (O_p) are iterated over in order of proximity. Each of the child objects assigned to a vantage point will have its distance to the query computed in turn. For range queries, there is an associated cut-off radius, d_{max} . As with the MTree, we may use a simple pruning rule using the stored parent distances to determine whether or not we even need to compute the distance between the query and the i 'th child object (Figure 14).

$$|d(Q, O_p) - d(O_i, O_p)| \leq d_{max}$$

If this holds then $d(Q, O_i)$ is calculated, and it is inserted into the NN priority queue. If the above condition does not hold, then it is checked whether $d(Q, O_p) < d(O_i, O_p)$. If so, as the children have been stored in sorted order, $d(O_{i+1}, O_p)$ must be greater than $d(O_i, O_p)$, so the remaining children in the vantage point may be skipped. For k -NN queries, after the first cluster has been processed, we may use the distance associated with the k 'th item of NN as d_{max} , and use this value to prune further searches.

Yet again, the vantage points covering radius may be used to remove entire clusters from the search. The minimal distance a cluster point could be from the query, d_{min} , is given by:

$$d_{min}(O_p) = \max \{0, d(Q, O_p) - r(O_p)\}$$

If the covering radius of a vantage point is greater than the distance between the query and the vantage point, then d_{min} will be 0 as the cluster could potentially contain an exact match. If d_{min} is less than

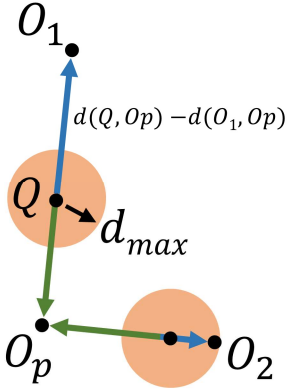


Figure 14: The same pruning rule demonstrated in Figure 12 may be used in the LC to prune children without computing their distance to the query

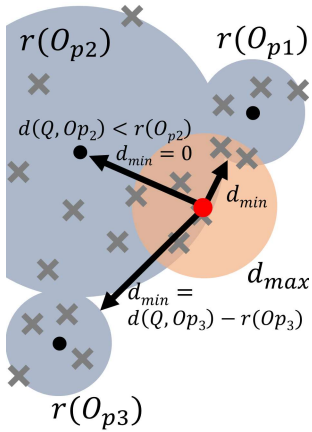


Figure 15: If d_{min} of a cluster is greater than the k 'th entry of NN, d_{max} , then this entire cluster may be pruned from the search.

d_{max} then the cluster could contain some objects which are similar to our query. Else if this d_{min} value exceeds the identified d_{max} value, then the cluster cannot have any objects relevant to our search and may be skipped (Figure 15).

Methods

Data

To use an MTree as a lookup tool, existing database records are required. Contemporary datasets of crystallographic files and experimental reports were acquired from existing APIs, custom built web scrapers, and the matminer repositories¹⁰. A summary of these datasets and the methods that these were obtained are given in Table 1. The ICSD client was created by myself to access the existing REST API, as the commercial provided solution for accessing the ICSD requires a further interface to be streamlined into a typical scientists python based tooling. This is accessible under an open-source licence at www.github.com/lrcfmd/ICSDClient, which has shown use by other materials science researchers on a range of projects, which has included gathering datasets for investigations into thermoelectric, solid state electrolyte and transparent conducting materials. Across the 3,005,510 records there are 1,065,165 unique compositions. This is used as the indexing set for each of the tests.

Table 1: The datasets collected and indexed with the EIMTree, their focus, and the method of dataset retrieval.

| Dataset Name | Number of Entries | Unique Compositions | Focus | Method of Obtaining Data |
|---|--------------------------|----------------------------|--|---------------------------------|
| Inorganic Crystal Structure Database ¹¹ | 200,809 | 123,356 | Inorganic Materials | Custom API |
| Cambridge Structural Database ¹² | 841,421 | 775,577 | Organic Materials | CCDC API |
| Materials Platform for Data Science (Pauling File ¹³) | 1,473,502 | 114,537 | Inorganic Materials | MPDS API |
| Materials Project 2018 ¹⁴ | 83,989 | 83,989 | DFT Computed Crystal Structures | Matminer API |
| Matminer Datasets ¹⁰ | 251,243 | 126,531 | A collection of experimental and calculated datasets for materials science | Matminer API |
| Pearsons Database (Pauling File ¹³) | 354,574 | 181,785 | Inorganic Materials | GUI interface |
| Total | 3,005,510 | 1,065,165 | | |

Tree Hyperparameter Selection

A range of hyperparameters were initially chosen to construct MTrees and Lists of Clusters, up to a maximum construction time of 12 hours. Each node of the MTree has a single internal routing object, and pointers to other nodes of the tree, the number of which is referred to as the internal node size. Increasing the node size will reduce the overall depth of the tree, as each node can route queries to a greater number of subtrees per level. However, a greater number of comparisons must be made at each level before deciding which nodes should be added to TQ. Node limits of 2, 3, 4, 5, 10, 15, 50, and 100 were chosen.

For the list of clusters, we may choose how many vantage points to select from the dataset at the start of the algorithm. A smaller number of vantage points, m , will mean that each vantage point will have a greater number of objects indexed beneath it. The average number of objects indexed to each vantage point is given by the ratio n/m , and for this investigation n/m ratios of 1024, 128, 64, 32, 16, and 8 were selected. As the LC architecture allows for parallelisation during indexing this was applied during testing, with a total of 16 processes spawned. To provide a measure of comparison, the brute force search times are also provided.

Results and Discussion

Indexing Times

Each data structure was indexed under each associated set of hyperparameters, terminating after 7 days. This time limit meant that some MTrees failed to index all 1,018,456 unique compositions. The total time taken to index the dataset for each run is given in Figure 16.

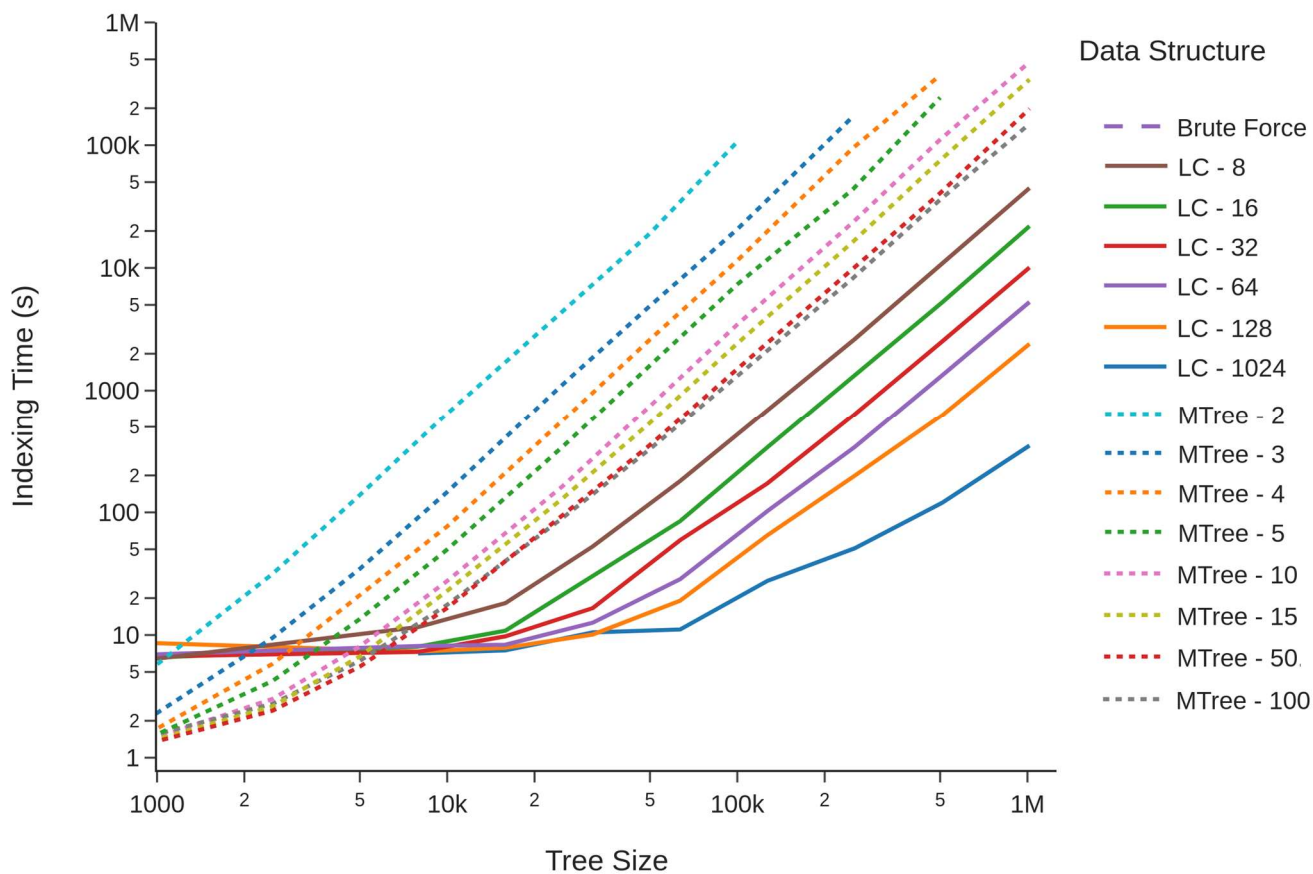


Figure 16: The total time taken to index each data structure. This is repeated for increasing numbers of compositions in the indexing set, thus increasing total tree size.

k -NN Search Times

After constructing each tree, 20 randomly selected compositions were selected from the dataset and a k -NN search performed, with results shown in Figures 17 and 18.

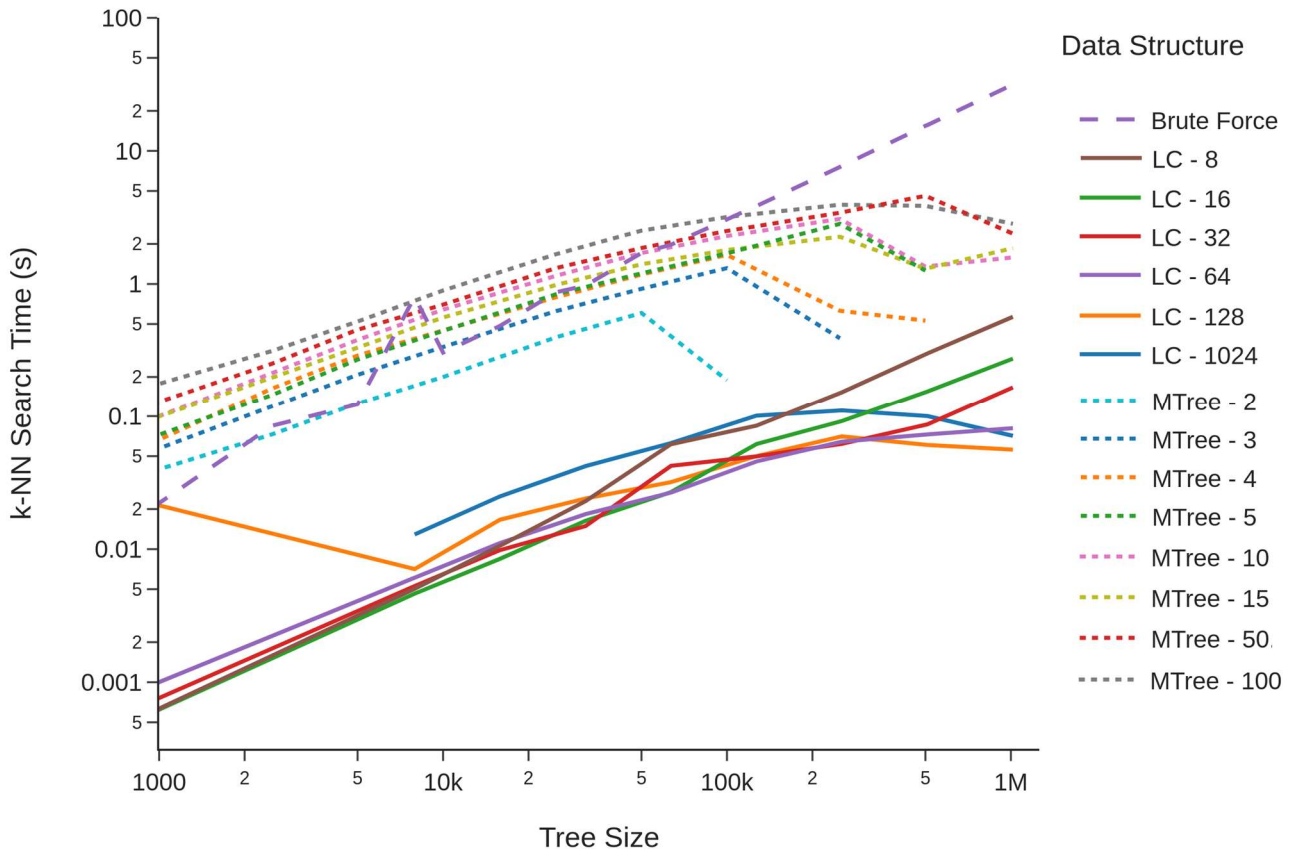


Figure 17: Average k -NN search times for 20 randomly selected compositions. This is repeated for increasing numbers of compositions in the indexing set, thus increasing total tree size. Different architectures and hyperparameters are given, with a brute force search overlaid.

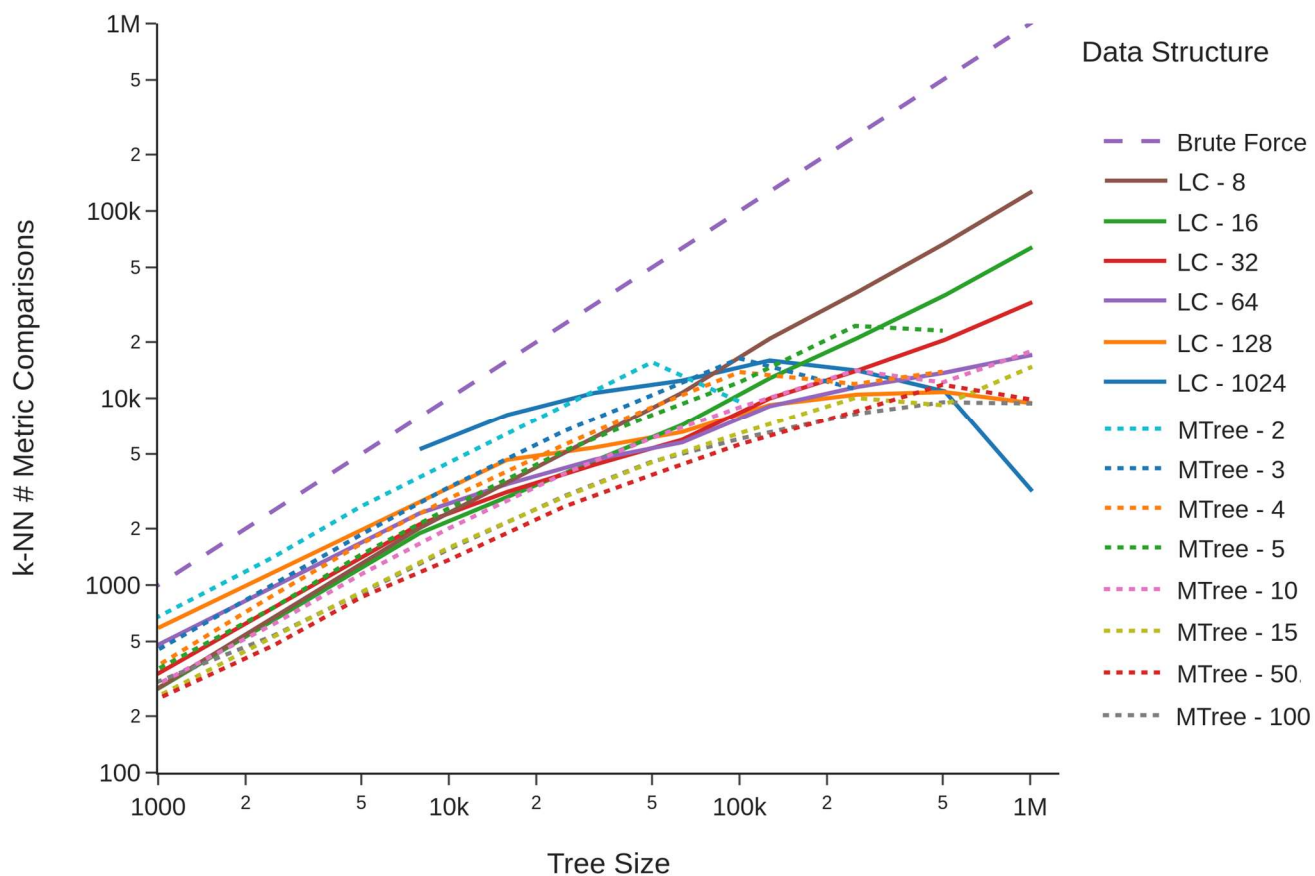


Figure 18: The total number of distance computations that are calculated when executing a k -NN search for 20 randomly selected compositions. This is repeated for increasing numbers of compositions in the indexing set, thus increasing total tree size. Different architectures and hyperparameters are given, with a brute force search overlaid.

Range Query Search Times

After constructing each tree, 20 randomly selected compositions were selected from the dataset and a range search performed, with results shown in Figures 19 and 20.

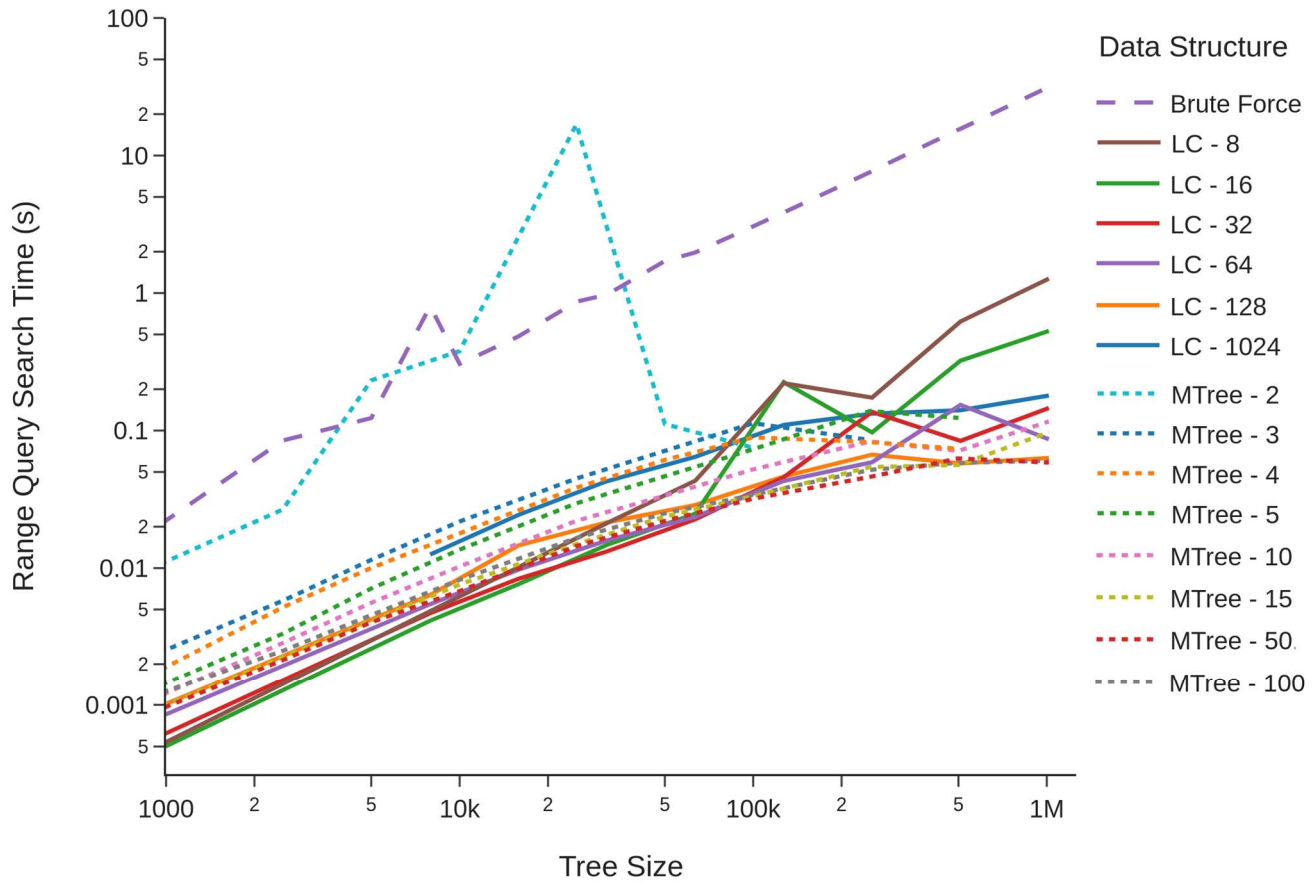


Figure 19: Average range query search times for 20 randomly selected compositions and a range cut-off radius of 1. This is repeated for increasing numbers of compositions in the indexing set, thus increasing total tree size. Different architectures and hyperparameters are given, with a brute force search overlaid.

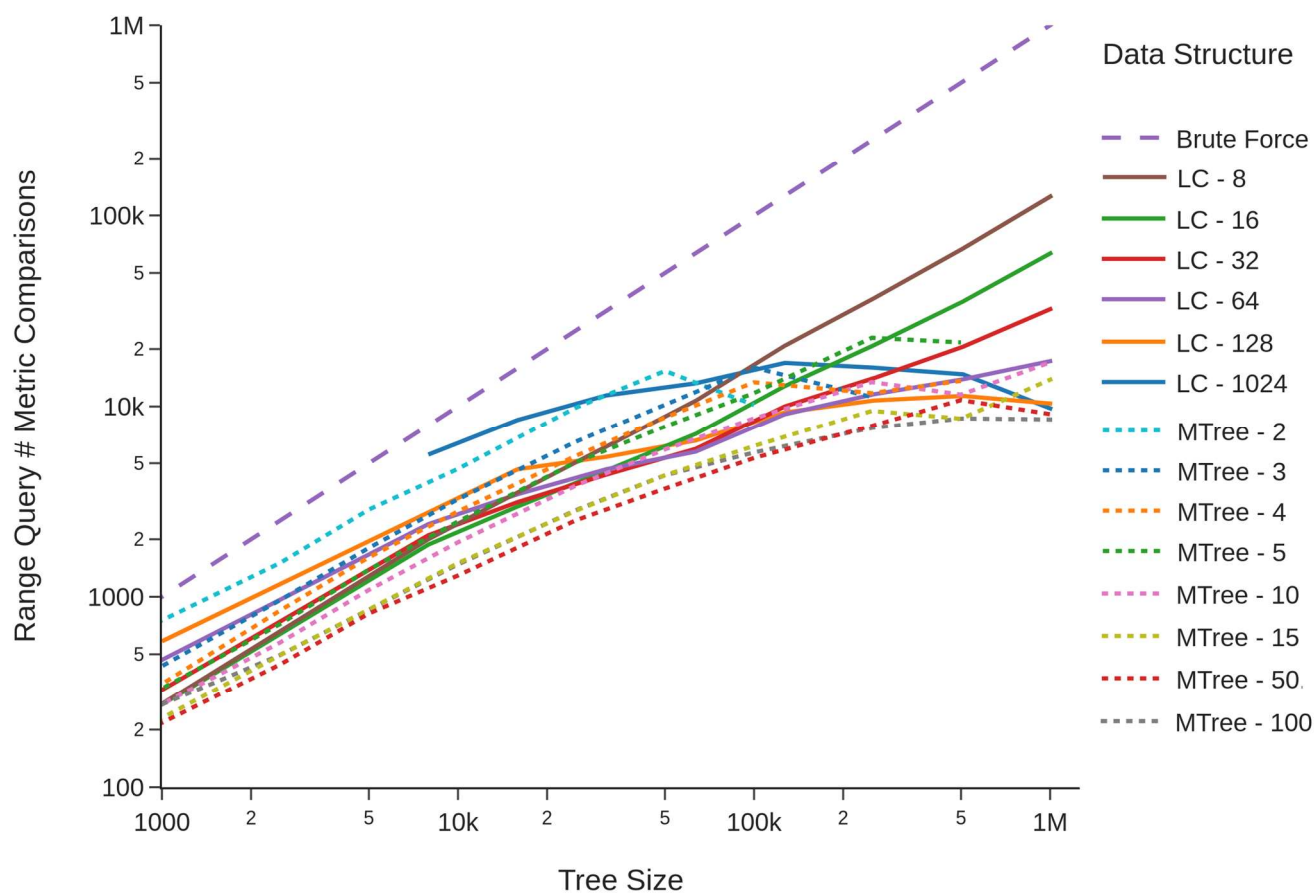


Figure 20: The total number of distance computations that are calculated when executing a range query search for 20 randomly selected compositions and a cut-off radius of 1. This is repeated for increasing numbers of compositions in the indexing set, thus increasing total tree size. Different architectures and hyperparameters are given, with a brute force search overlaid.

Whilst each of the two data structures execute similar numbers of metric queries during the k -NN and range query searches, the time taken for k -NN searches is significantly larger for the MTree than the LC. After analysis it was found that the MTree carries forward many more traversals of the tree when it is executing k -NN searches, with the TQ list being appended to many times. As memory management in python is typically slow, it is suspected that this causes significant overhead to the search process. By comparison the LC has a much simpler architecture.

It may seem counter intuitive that the LC gives comparable, and improved, performance to other methods. In this instance, with 9,675 vantage points, the LC root has an analogous node limit that is significantly higher than tested, or even could be practically supported, by the MTree. Constructing such a large distance matrix each time a node requires promotion in the MTree would be prohibitively time and memory intensive. As the LC is technically a very broad tree, it does not appear to be at all balanced. Executing 9,765 similarity queries as an initial step seems to go against the core premise of tree indexing schemes. We should not need to carry forward thousands of computations. We should be able to do this in logarithmic time.

The difficulty of this problem arises in the overlapping covering radii, as in practice these force us to explore many regions which may not be similar to our queries to ensure we have not missed an object. By contrast, the high degree of fragmentation that the LC pivots provide, means that despite the heavy initial search cost, the large number of vantage points means that these will tend to follow the natural distribution of the space. Specific regions of the space can be located much faster as a consequence.

Given the reasonable indexing times, and overall favourable performance for k -NN searches, an LC with a n/m ratio of 1024 was chosen to be implemented for the live interface. This means that each k -NN search will perform a minimum of 9,765 queries to each of the pivots before iterating through each of their children.

Hosted Search Interface

A web interface was developed to serve ElMTree queries for ElMD similarity searches (<https://lmds.liverpool.ac.uk/ElMTree>), using the python library flask to process queries and render HTML. Simple networked applications can be shared insecurely to colleagues over internal networks with no additional setup. However, sharing applications with the wider research community in a reliable, secure, and scalable manner requires additional infrastructure. Traditionally, high availability applications run on servers in data centres, with best practice being to give each logical component (databases, applications, networking etc.) a dedicated server. Advances in the reliability of electronics, and processor technologies, particularly in CPU virtualisation, mean that standard practice is now to emulate each server inside a virtual machine (VM) with many VMs running on a single server, or a cloud of servers. Commercial cloud computing platforms require specific technical expertise to operate, and estimating the long term running costs is difficult. Due to the availability of an on-premises IT support team to assist in setup and running operations, an on-premises cloud solution was employed (Figure 21).

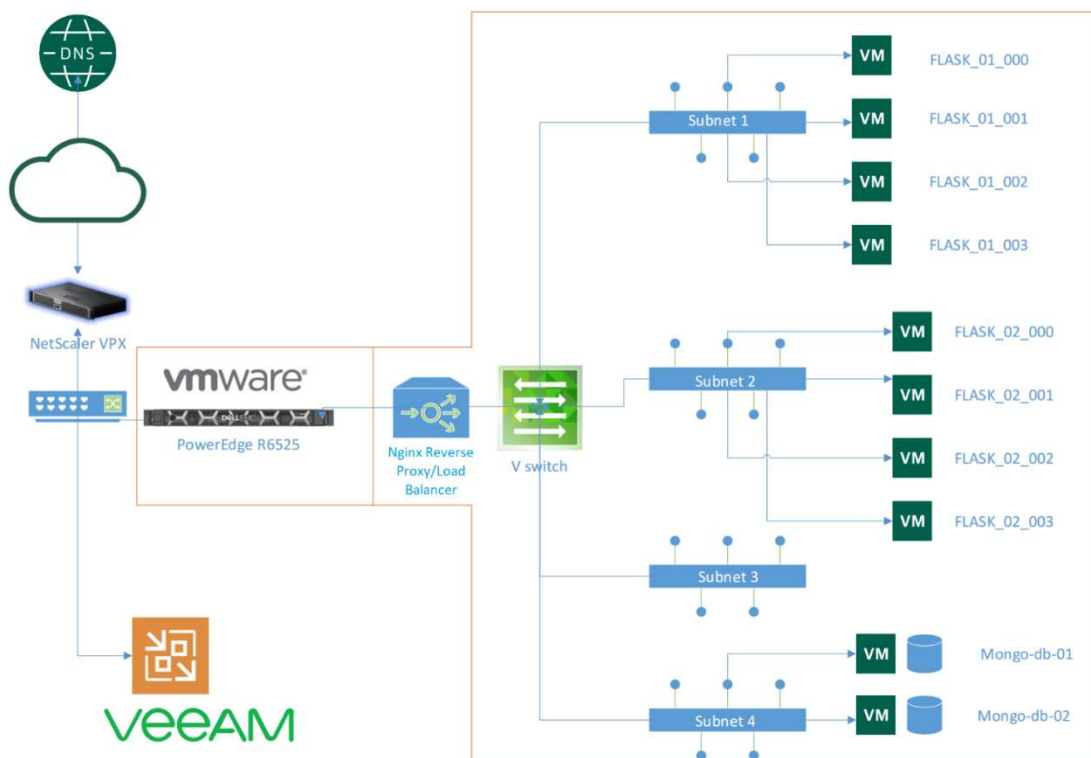




Figure 21: The on-premises cloud architecture that has been deployed as part of the Liverpool Materials Discovery Server. External traffic is routed through the university NetScaler, all VMs run inside a PowerEdge R6525 running VMWare vSphere. Requests are directed to the correct service through the Nginx reverse proxy and load balancer, with each application running in an isolated VM.

VMWare vSphere was installed on a Dell PowerEdge R6525 server to serve as the base hypervisor for all VMs. Each flask application is hosted on a separate ubuntu VM using the unicorn library to keep multiple processes of the application running, and to recover each process if it crashes. HTTP requests are routed internally and externally via an Nginx reverse proxy VM. The Nginx reverse proxy allows a single secure socket layer (SSL) certificate to be shared across all applications, which lets external users verify that network traffic has not been intercepted or modified by a malicious party. The reverse proxy may also be used as a load balancer to distribute requests amongst application VMs evenly and as a rate limiter to prevent targeted botnet attacks from crashing the site.

If the network traffic of a popular application increases significantly, this will affect the performance of that application for all users. The computing capacity of an application can be expanded by cloning the associated VM, updating its IP address, and adding this to the list of routable IP addresses in the reverse proxy. Additionally, each VM has its state saved to the on-premises backup farm daily with the Veeam software, further guaranteeing recoverability in case of major hardware failure. To ensure applications have a consistent and high quality appearance across devices, the software development firm CloudPloys were contracted to design and code front end styling, Figure 22.



UNIVERSITY OF
LIVERPOOL



ELMTree

Searched 1,065,165 unique compositions across 3,005,510 records for Na0.5 Cl0.5 in 0.419s

| Composition | Database: Matched IDs | Distance |
|-----------------------|---|----------|
| Na0.5 Cl0.5 | <ul style="list-style-type: none"> • icsd: icsd_018189, icsd_028948, icsd_041411, icsd_041439, icsd_043434, icsd_052232, icsd_052233, icsd_053815, icsd_053816, icsd_060280, icsd_061662, icsd_100633, icsd_165592, icsd_169462, icsd_181148, icsd_197693, icsd_240598, icsd_240599, icsd_240600, icsd_240601, icsd_240602, icsd_240603, icsd_240604, icsd_622368, icsd_622369 • boltztrap_mp: mp-22862 • brgoch_superhard_training: mp-22862 • elastic_tensor_2015: mp-22862 • expt_formation_enthalpy_kingsbury: 514 • expt_gap: 1462 • jarvis_dft_3d: mp-22862 • jarvis_ml_dft_training: mp-22862 • matbench_dielectric: 1534 • matbench_expt_is_metal: 3028 • matbench_log_gvrh: 1635 • matbench_log_kvrrh: 1635 • mp_20181018: mp-22851 • mp_e_form: 88268 • pearsons: True | 0.0 |
| K0.05 Na0.45 Cl0.5 | <ul style="list-style-type: none"> • icsd: icsd_028947, icsd_240594, icsd_240595, icsd_240596, icsd_240597 | 0.05 |
| Na0.5 Br0.051 Cl0.449 | <ul style="list-style-type: none"> • icsd: icsd_061663 | 0.051 |
| K0.088 Na0.412 Cl0.5 | <ul style="list-style-type: none"> • icsd: icsd_028946 | 0.088 |
| Na0.5 Br0.1 Cl0.4 | <ul style="list-style-type: none"> • mp_e_form: 84212 | 0.1 |

Figure 22: A screenshot of the ELMTree search interface, demonstrating a search for the composition NaCl.

Stress Testing

To determine the capacity of this server to process requests, a single process Flask application was deployed. This was stress tested by selecting compositions at random and directing increasing numbers of database queries at the server using the Locust library. This starts at 1 request per second, increasing by increments of 5 requests every second until the server is being requested to query the database 750 times a second, Figure 23.



Figure 23: Stress test results running the ElMTree application as a single process. Response times begin to rise with more than 100 concurrent users, with the majority of responses failing when there are more than 500 users.

Response times are very reasonable when processing fewer than 100 requests a second, returning requests in under a second on average. As the load increases, these response times increase to over 10 seconds with 500 concurrent users. Eventually, after approximately 520 requests a second, the server responses become much shorter on average, as a greater proportion of the responses fail due to server overload. This can be alleviated by scaling the capabilities of the application.

The hosted architecture allows this to be enacted by changing two configuration files, one to increase the VM capacity, and the second to increase the number of worker processes. The same stress test is repeated, first with 4 cores, 40 GB of RAM, and 2 worker processes, and again with 8 cores, 80 GB of RAM, and 4 worker processes, with the results shown in Figures 24 and 25 respectively.

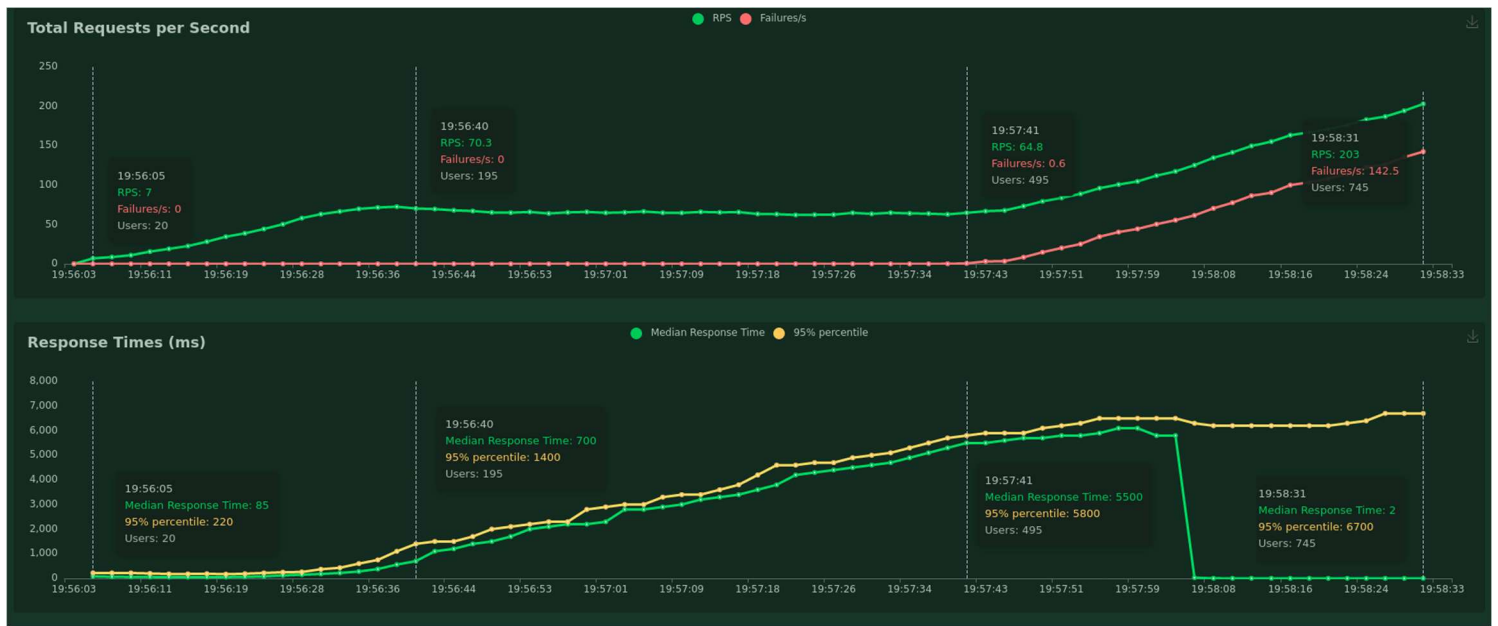


Figure 24: The stress test results of the ElMTree application when increasing the number of worker processes to 2.

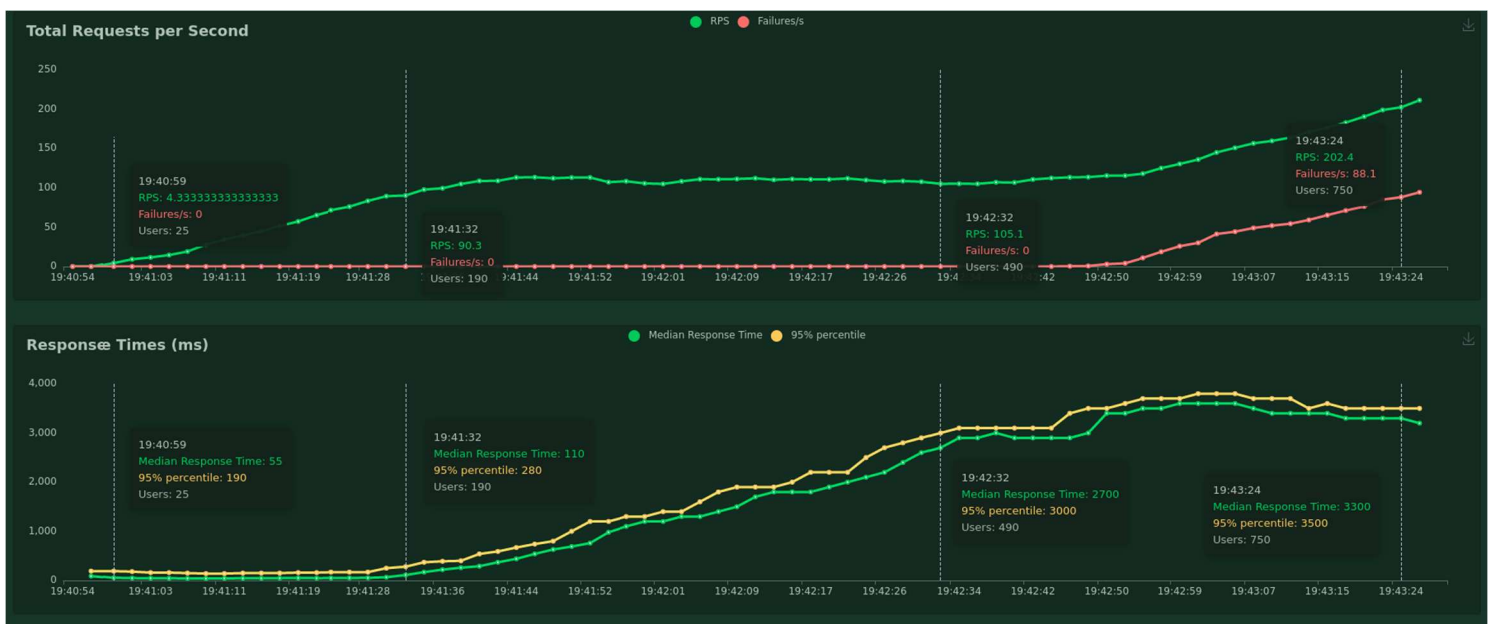


Figure 25: The stress test results of the ElMTree application when increasing the number of worker processes to 4.

As expected, scaling up the architecture allows the application to serve a greater number of requests, with a shorter response time on average. Given that in typical usage we would not expect more than 1 request a second, for general purposes a single process has been used for the live EIMTree. However, should performance degrade with use, or an expected period of heavy activity is forecast, these stress tests demonstrate that the architecture should remain usable by scaling the available resources.

Conclusions

In many investigations, the earliest step is the observation of a trend. Metric indexing trees allow us to organise collections of data in a logical fashion, enabling the swift execution of similarity queries. This comes with obvious advantages, as the retrieval of data is the core reason for storing data in the first place. Search interfaces remain the established method of interacting with databases for non-technical users, and improvements in search speed and quality provide immediate benefits to researchers in the field.

As any valid metric may be used to index a set of objects into the List of Clusters, further work could investigate the application of more focussed materials metrics. Chemical similarity is a reasonable starting point for the chemist to judge the similarity of two materials. But, we know this crude identifier does not capture the intricacies of a compounds properties. The use of structural metrics can distinguish materials that many chemists would call chemically similar, such as the polymorphs of a composition. Further, if we know we are searching for materials with a particular property, such as ionic conductivity, then we may wish to search for those structural motifs which are known to correlate with the property of interest.

The next chapter outlines a descriptor which was developed for this purpose, Percifter. In its current format the construction remains too computationally costly to execute large searches, but it has

demonstrated itself a useful tool for surveying the motifs present across the reported families of conductors in the dataset gathered in chapter 3. Nevertheless, we believe the flexibility afforded by this architecture enables future investigations into applying bespoke metrics for focussed searches into potential candidate materials from previously reported datasets.

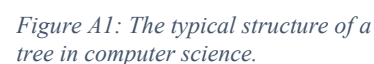
Noise is the life partner of every observation, inescapably introduced both in experimental error and the generation losses that are introduced through the digitising processes. The ready availability of both experimentally derived and calculated data in materials science can present an overwhelming tide of information, tangling observable patterns into a sea of text. The use of these tools allows us to consider queries against restricted and chemically sensible subsets of the reported compounds, instead of being presented against a backdrop of irrelevant data. This promotes the formation of new theories to explain trends which may have been obscured by the noise introduced by unrelated materials. Filtering and refining large datasets remains the preliminary step for many materials science investigations, and we believe the architecture presented here and deployed at <https://lmds.liverpool.ac.uk/ELMTree>, has further utility as a part of this process.

References

1. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms, fourth edition*. (MIT Press, 2022).
2. Bentley, J. L. Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**, 509–517 (1975).
3. Samat, H. *Foundations of Multidimensional and Metric Data Structures*.
4. Chen, L. *et al.* Indexing Metric Spaces for Exact Similarity Search. *ACM Comput. Surv.* (2022) doi:10.1145/3534963.
5. Ciaccia, P., Patella, M., Rabitti, F. & Zezula, P. Indexing Metric Spaces with M-tree. 20.
6. Sadit Tellez, E. & Chávez, E. The List of Clusters Revisited. in *Pattern Recognition* (eds. Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., Olvera López, J. A. & Boyer, K. L.) 187–196 (Springer, 2012). doi:10.1007/978-3-642-31149-9_19.

- # Appendix

As each child contains pointers to further sub-children, we may follow paths of many nodes from the root until we encounter the terminal leaf nodes, which do not contain further pointers to objects. The space may be split into regions called subtrees by considering an internal node and its descendants, with the subtree of an node object O denoted $T(O)$. We may disregard large



portions of the search space when executing similarity queries by ignoring whole subtrees of the search tree.

The binary search tree is a common introduction ¹, as it is simple to elucidate. This data structure is only applicable on sets of objects with well-defined ordering functions, which is to say the $<$, and $>$ operations may be consistently applied between any two objects in the set. Each internal node of a binary tree will contain no more than two pointers to other nodes. At each branch, the left edge will point to a child which is less than the node object, and the right branch points to a child which is greater than the node object.

During searches, we take our query object and at each level ascertain whether it is greater or less than the current node. Certain objects, such as binary strings, can be retrieved in constant time using this data structure. For example, using the letters of the alphabet encoded by their Morse code representations, with the ordering rule that dot is less than dash, we get an efficient method of retrieving the associated letter of the alphabet when decoding Morse code messages. The lexicographic ordering of the Morse letter encodings may be read off from left to right as the tree is descended (Figure A2).

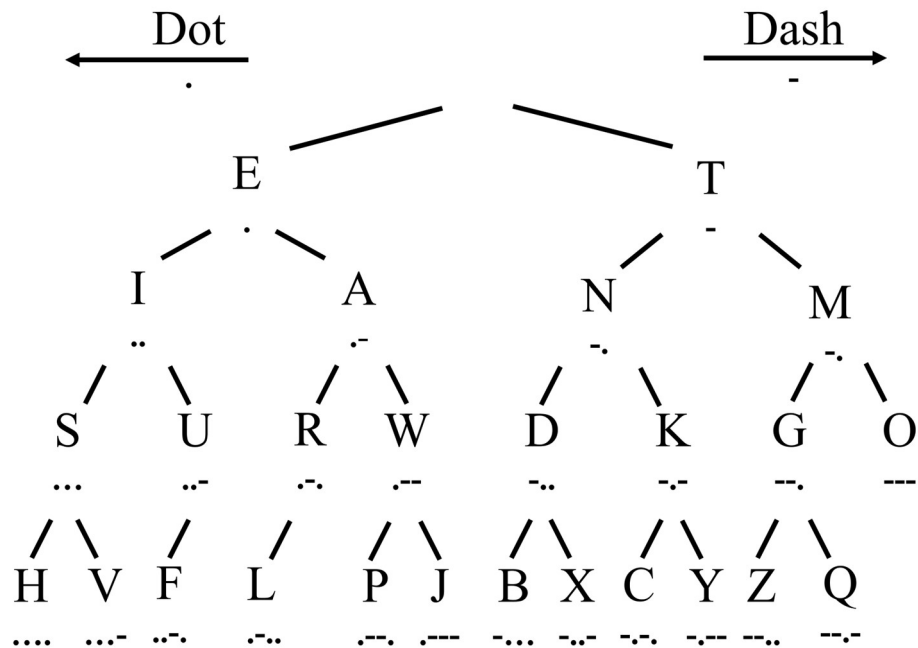


Figure A2: The letters of the alphabet indexed into a search tree following the lexicographic ordering of their associated Morse code.

Constructing binary search trees can be simple. The first object of an initially unordered set is made to be the root object. Then we iterate over the set, routing each new object to the best leaf to insert a new leaf node containing the object. If an object is less than the root object, descend the tree to the left edge and examine the object that is being pointed to, else descend to the right. This process is repeated on each child node in turn, until an empty node pointer is encountered, in which case we assign the new indexing object as the child of the node that we are currently at.

Unfortunately, this approach can lead to poor search performance, no better than that of an array in the worst case. Let us demonstrate this by using numbers from the interval $[1, 6] \in \mathbb{Z}$ as our indexing set. If these are inserted in the order $[4, 2, 5, 1, 3, 6]$, then the tree will be rooted at 4, with 2 and 5 as child objects. 1 and 3 will be children of 2, and 6 will be a child of 5. If we wanted to run a retrieval query for the number 6, we would first check it against 4, then 5, before encountering 6. However, if the objects are inserted in a sorted fashion, then the root will be equal to 1, it will have a single child, 2, which will have a single child, 3, and so on and so forth. This means our retrieval query for 6 will need to compare against every single object in the set before encountering the number 6. As each of the edges point to the right, we would say this is a highly unbalanced tree. This makes the tree much taller and thinner than we would like, which will extend the time taken for similarity queries to execute.

In a balanced binary tree, the root of the tree will contain the object that is the median of the indexing set, with each internal node being located the median object of all its' children and sub-children. If we wish to enforce a balanced tree, we must execute a tree balancing algorithm, which will re-assign the trees node pointers until the tree is balanced. This operation will often take significantly more operations than a simple insertion, which may make the procedure unsuitable to carry forward on large trees. Furthermore, in this example where we have a uniformly

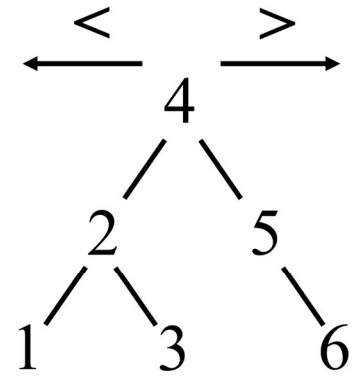


Figure A3: A balanced binary search tree for the integers 1 through 6.

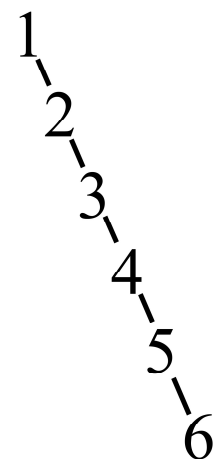


Figure A4: An imbalanced binary search tree for the integers 1 through 6.

distributed set of objects that need indexing, it is clear that the unbalanced tree is an inferior arrangement. However, for highly irregular distributions, balanced trees have the potential to reduce the overall efficiency of the search. Distant or sparse regions of the search space may be forced into local proximity to one another in the search tree despite sharing very little similarity to one another, to ensure the tree maintains balance. This will force searches to compute a great number of redundant comparisons.

The binary tree is suitable for 1-dimensional numeric data as this space has a consistent ordering attached to it. Unfortunately, the high dimensional vectors we use to represent materials do not possess a canonical ordering. Even objects represented by 2-dimensional vectors, such as points on the plane, cannot be indexed in such a manner. The classical method of indexing vectors in k -dimensional space, $k > 1$, is a generalisation of the binary tree process, the k -d tree ².

In the k -d tree, subsequent axes of the vector space are considered to split the space in two as we descend the tree. From the root, this means that the children of the left child's first element will be less than the first element of the root vector, and items in the right child being greater. This only considers the first axis of the vector space, and so at the next level, the second element of the vector is used to route objects left and right instead. This will continue to split on each of the axes of the vector space in turn until the $k+1$ level of the tree, at which point the first axis will be under consideration once again. This allows the processing of high dimensional vector spaces, and can be fast in practice.

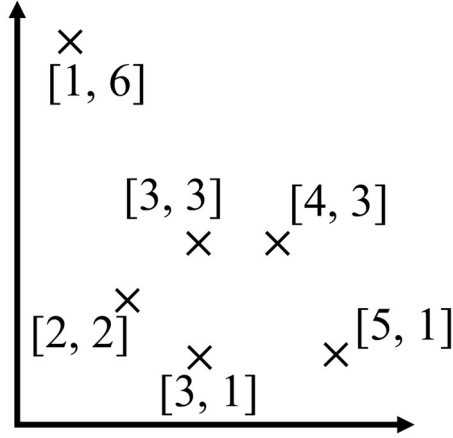


Figure A5: A set of six points in 2-dimensional space

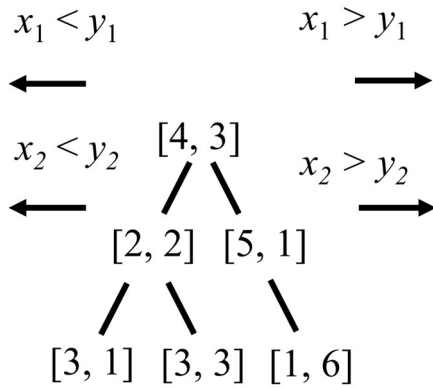


Figure A6: The points of Figure A5 when indexed with a k -d tree.

Percifter: A topological descriptor for periodic crystal structures

This is an original piece of work, conceived by myself and Matthew Dyer that is yet to be prepared for publication. We are very thankful for Matthew Bright's continued input and insight throughout this project, and to John Claridge, Vitaliy Kurlin, and Matt Rosseinsky for their support and guidance.

Introduction

Defining a material through its composition is advantageous due to the time consuming techniques required to characterize a crystal structure experimentally or computationally. Nevertheless, the physical arrangement of atoms in a crystal play an integral role in determining a materials physical properties, and this additional information often gives us the ability to predict these properties through physical and statistical modelling to a high degree of accuracy ¹⁻⁴. Mapping perovskites by their ionic radii was one of the earliest investigations to classifying materials based on their structural characteristics ⁵. More recent investigations seek to encode structure through hand crafted physical descriptors which have been reasoned to correlate with a property of interest ⁶. This has a level of interpretability, as we can correlate known physical features with observed target properties, and focus future investigations to exploit these observed characteristics.

Unfortunately this approach is prone to human bias, and it may not arrive at the optimal combination of structural features for the strongest correlation with the target property. In this work we present an automated characterisation procedure that provides an interpretable representation of a crystal structure that may be used with ML models.

It is often found that more complex structural descriptors correlate with better performance at predicting physical properties with ML models, at the expense of human interpretability. A widely used

representation is the smooth overlap of atomic positions (SOAP) descriptor ⁷. Here a function is generated for each atomic position to represent the contributing overlap of nearby atomic sites, attaining permutation invariance by summing these together to give the final representation of the crystal. This is calculated by first placing a three-dimensional Gaussian distribution at all atomic sites in the structure. One-dimensional rotationally invariant smeared atomic positions are obtained by integrating Laplace spherical harmonic functions (Figure 1) by the volume of the sphere at increasing radial cut-off values for increasing harmonics, up to n_{\max} . A fixed length representation is obtained by summing each of the harmonics and calculating the power spectrum of each site and combining these to give an n -dimensional vector. This is commonly generated with the DScibe package ⁸, and can be used to compare structural similarity between two crystals directly using the Euclidean metric or the EMD. The SOAP descriptor is an invariant representation which demonstrably correlates with physical properties, and has been well applied in the wider literature ^{9,10}. One observation of the SOAP descriptor is that it gives equal weighting to the power density at all radial values. It may be that only certain interactions between elements in the material are truly correlated with the property of interest, and the addition of all other interactions reduces the signal to noise ratio of the relationship we are trying to predict.

Recent graph neural network (GNN) ¹¹ architectures have been proposed to alleviate this, with crystal graph convolutional neural network (CGCNN) ¹² being one of the more popular implementations. An elemental representation is generated for each site in the structure, which is iteratively updated as it passes through the initial graph pooling layers with information about each of its neighbouring atomic site vectors. The final representations of each atomic site are summed together, and this fixed length vector is passed into a feed-forward neural network to make a prediction and calculate the error against the true value. The weighting that is given by each site to its specific neighbours

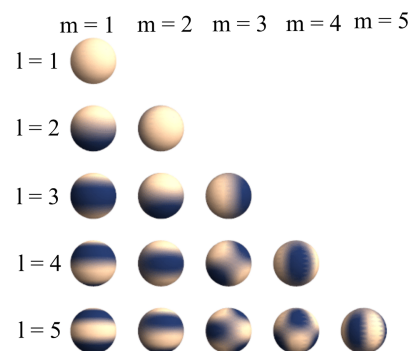


Figure 1: The spherical harmonics.
en.wikipedia.org/wiki/Spherical_harmonics

before adding the neighbours vector to the atoms vector is a parameter that is learnt during training. It is hoped that interactions which are observed to give the best increase in predictive performance are identified through gradient descent and given a greater weighting in the trained network. The final trained graph pooling network should focus on interactions between neighbouring sites which are most correlative with the target property. A hyperparameter that must be established for ionic crystals prior to training is the edge set that will be used to consider two atomic sites as being neighbours.

Expressing the topology of a crystal involves adding edges between atoms, constructing the corresponding simplicial complex, and computing the Betti numbers of this complex. These terms will be discussed in greater detail later, but one fundamental parameter this approach requires, is a cut-off radius to be defined to construct edges between each atomic site ¹³. CGCNN, uses a cut-off radius of 8Å as it is assumed that no inter-atomic interactions can take place at a distance greater than this. This same technique is used in further developments which use GNNs ^{14,15} to construct message passing networks. This method of representing a crystal can consistently describe the topology, and is state of the art in supervised learning, but has identified limitations.

The Weisfeiler-Lehmen isomorphism test ¹⁶ allows us to assess whether two graphs are the same (isomorphic). Due to the method of vector pooling employed by GNNs, two structures with dissimilar atoms but similar connectivity graphs could be indistinguishable to the neural network. The Weisfeiler-Lehmen isomorphism test could however demonstrate the two are in fact dissimilar, which is an underpinning restriction of the GNN message passing paradigm. It has further been remarked that GNNs will implicitly focus on short range interactions, as node features do not typically propagate to extended neighbourhoods. More recent approaches aim to remedy this shortcoming by using provably invariant representations of the structure which possess

“better” internal symmetries. $E(3)$ equivariant networks extend the use of fixed spherical harmonic functions by learning the parameters of such functions through back-propagation, but requires training data to learn the representations.

By contrast, the recently published pointwise distance distribution (PDD) ¹⁷ is a simple to compute representation which can fully distinguish any two periodic structures. This involves taking the distances from each atomic site to the k -nearest other atomic sites in the structure. These distances are then sorted to form a vector for each site of length k . These vectors stored in sets of tuples with the associated ratio of the vector across all sites. We shall explore the PDD representation under the EMD metric in further sections.

In this chapter we layout a separate approach based on the advances made in topological data analysis ¹⁸, in particular the field of persistent homology ^{19,20}. Several prerequisites are presented at an introductory level, followed by a discussion of persistent homology in the finite case, before moving on to the issues introduced by periodic boundary conditions and our proposed augmentations to the theory to create stable crystal structure descriptors using the Percifter package. An analysis of this descriptor applied to simple datasets is presented with some examples integrating the associated metric of similarity into downstream tasks.

Prerequisites

An overview to the field of topological data analysis is beyond the scope of this chapter. Fortunately, the core algorithm we will take from this toolkit, persistent homology (PH), may be explained succinctly. As with neural networks, a reasonable understanding of the process, outcomes, and benefits can be ascertained by a materials scientist without prior specialist domain knowledge, although several fundamental definitions need to be introduced first. We give a very general definition of homology before focussing on the computational realisation used in this work, which

may be described using the language of vector spaces and linear transforms. A reader familiar with the topological concepts of fundamental groups and covering spaces can attain a more thorough overview in the textbooks by Hatcher ²¹ or Munkres ²². Computational Topology by Herbert Edelsbrunner ²⁴ is the best introductory textbook to persistent homology.

As opposed to supervised learning where a prediction for a vector is generated based on statistical exposure to samples of a training set, PH is an unsupervised technique which takes a (potentially high dimensional) pointcloud of points in row vector form, and returns a quantified representation of the topological voids that form between these points as we enlarge each 0-dimensional point into a solid sphere of expanding diameter, loosely termed as the homology of the space. The output is a list of pairs of real numbers, [*birth*, *death*), with each pair representing a single topological feature, the values being the spherical diameters at which features form and cease to exist. As long sequences of numbers are not immediately interpretable to the human researcher, these are typically plotted as a graphical representation, either on a persistence barcode or a persistence diagram.

Betti Numbers

What are topological features? Most introductory texts will begin by discussing how coffee cups and donuts can be considered the same objects when made with continuously deformable pieces of rubber. Continuously deformable here means that there exists a map from each point on the surface of one space (the coffee mug) to the second space (the torus), which does not tear or glue together any sections of the surface. These need not be linear transformations, and we can stretch and squeeze the space however we please. However, should we puncture a hole, or join the edges of an existing hole together to make it whole, then we would say that the space has become topologically distinguishable from the original. This space may be similar, but we cannot have a single

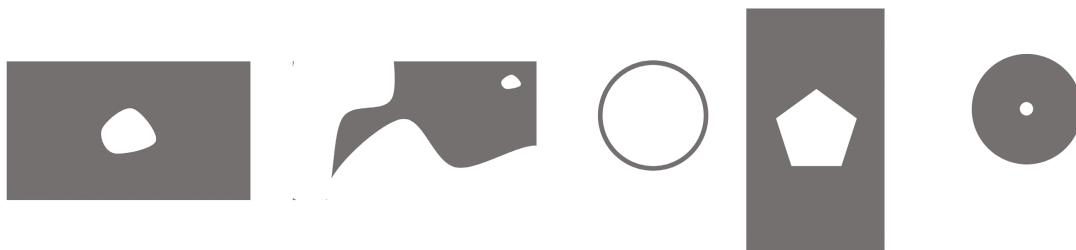


Figure 2: The first sheet of rubber may be continuously deformed into the second by stretching and squeezing the sheet. Each of these shapes may be deformed from the first, and are thus topologically equivalent.

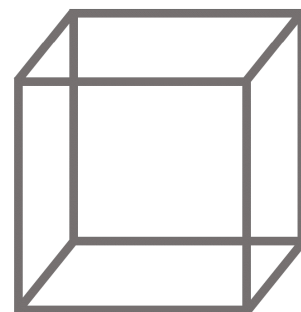
mapping function from one space to the other without introducing additional constraints on non-continuity.

Homology gives the language to describe these similarities. A homological equivalence, or homomorphism, means that two spaces may be continuously transformed into one another without introducing breaks or joins via a function $f: X \rightarrow Y$ that has a continuous inverse $f^{-1}: Y \rightarrow X$. This is easiest to visualise in dimension 1, and H_1 contains objects which are deformable to S^1 , which is equivalent to the standard circle made from a 1-dimensional shape, the line. Going down a homology dimension examines objects in $H_0(X)$, which describes the number of disjoint connected components in a space. Correspondingly, the space S^0 is a 0-dimensional point, and points, lines, planes, and solid 3-dimensional shapes (not hollow ones) are all topologically equivalent and mappable to S^0 .

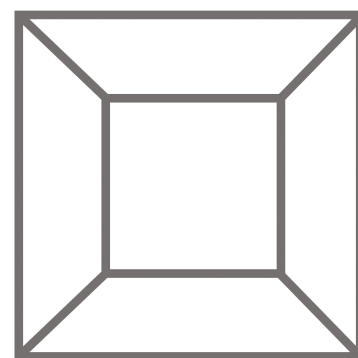
We will use a tool called the Betti number to measure this similarity between spaces. The k 'th Betti number of a space describes the number of k -dimensional holes in a space. More formally these count the number of generators in each homology class, $H_k(X)$. Whilst higher dimensional homology classes can be, and often are, investigated, in this study we restrict ourselves to H_0 and H_1 .

Let us demonstrate this on the sheets of rubber in Figure 2. B_1 , measures the number of holes that a space contains in H_1 , and a sheet of rubber with a single hole in it would be said to have $B_1=1$. Through continuous deformations, this sheet of rubber could be contracted into a circle, which is also a topological space with a

(a)



(b)



(c)

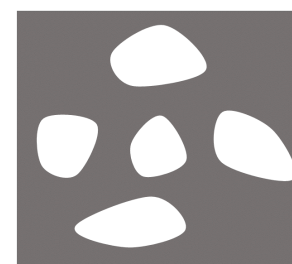
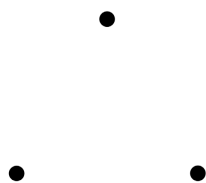


Figure 3: The wireframe cube (a) may be deformed to (b) by stretching out a rear corner and flattening the shape. By enlarging each of the edges, we see that this is topologically equivalent to a sheet of rubber with five holes in it.

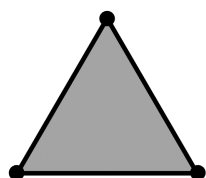
(a)



$$B_0 = 3$$

$$B_1 = 0$$

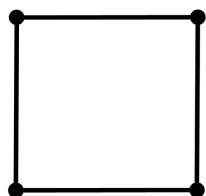
(b)



$$B_0 = 1$$

$$B_1 = 0$$

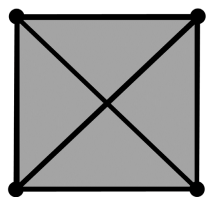
(c)



$$B_0 = 1$$

$$B_1 = 1$$

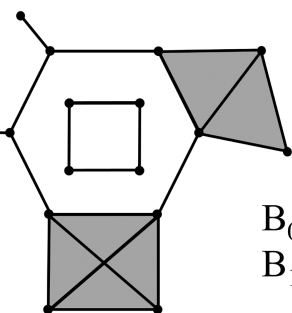
(d)



$$B_0 = 1$$

$$B_1 = 0$$

(e)



$$B_0 = 2$$

$$B_1 = 2$$

Figure 4: Five topological spaces and their associated Betti numbers. Three disconnected components, or points (a) form no cycles. Connecting these with edges forms a single filled connected component (b) with no cycles. Four points connected by four edges will form a single component with a single cycle (c). By fully connecting four points, we say a space is simply connected, which is equivalent to the filled tetrahedron (d) which has no cycles. Each of the simply connected regions of (e) may be contracted to give two disconnected cycles.

single hole. In fact, the five sheets of rubber above are all equivalent when considered under B_1 similarity. If we were to consider all five of these sheets together as a single topological space, this would be B_1 equivalent to the space consisting of five circles on the plane, as both spaces have $B_1 = 5$.

This does not tell us whether any of these sheets of rubber are connected to one another, and with just this information, this combined space is topologically indistinguishable to the wireframe of a cube, Figure 3. This may initially seem an erroneous statement, as we could expect the cube to be B_1 equivalent to the space of 6 circles glued by their edges. By flattening the wireframe of the cube onto the plane and expanding each of the edges out, we see that this is in fact a single sheet of rubber with five holes.

To describe the connectivity of the space, the Betti number B_0 may be used to count the number of connected components contained in a space. Looking at each of our five sheets of rubber as separate topological spaces, we can see that they each consist of a single connected component, therefore $B_0 = 1$ each time. Considering all of these as a joint space, this consists of five sheets of rubber which are disconnected from one another and thus $B_0 = 5$. By comparison, the wireframe cube is a single sheet of rubber with five holes punctured in it, but the cube has $B_0 = 1$, which allows it to be topologically distinguished from five disconnected sheets of rubber with holes in. Further examples are given in Figure 4.

Homology

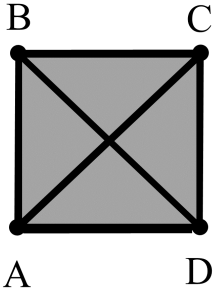
Before providing further examples of homology let us describe the mathematical procedure. Homology is the process of assigning mathematical spaces, such as topological spaces, to sequences of algebraic objects, such as groups or vector spaces, then using these group representations to judge when two spaces are “the same”, *homo-*. Groups and their group operations obey well studied rules of algebra, which provide a familiar setting for mathematicians to reason about abstract spaces. To describe a space, X , we use a

sequence of groups $H_k(X)$, to describe each of the k -dimensional voids contained in the space. If a group has an invariant representation, then we may say two groups are equivalent if the two invariants are equivalent. Two spaces are thus equivalent if all of their homology group invariants are equivalent.

In topological spaces, the group invariant that we have discussed to describe the k -th homology class $H_k(X)$, is the Betti number, B_n , which defines the rank (number of linearly independent generators) of the abelian group that represents the k -dimensional holes of a space. Each of the groups of a given topological space will have a characteristic Betti number. If two spaces carry identical Betti numbers, then we say the groups are homomorphic, and we know a homomorphism, or a continuous mapping, must exist between the two topological spaces, making them topologically equivalent.

For the 2-dimensional examples using holes in a sheet of rubber, no continuous deformations exist which take these sheets to spaces which are equivalent to the hollow sphere, S^2 . No matter how this space is stretched, there will never be a 3-dimensional void with an associated generator in H_2 unless parts of the space are glued together. This means that the associated Betti number B_2 will be equal to 0. The next homology class, $H_3(X)$ will have B_3 equal to 0, and so on and so forth for all higher homology classes. One of the core goals of homology is identifying the underlying groups which differentiate spaces. For persistent homology we shall represent these groups using a matrix containing only ones and zeros, marking the presence and absence of features of each space. These matrices may be reduced with Gaussian elimination under modulo 2 arithmetic (discussed below). Spaces that are topologically indistinguishable will reduce to the same set of matrices as part of the persistent homology process. Topologically similar spaces should reduce to a similar set of matrices.

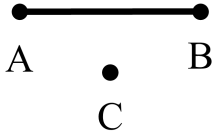
(a)



$$X = \{\sigma_1^3\}$$

$$= \{ABCD\}$$

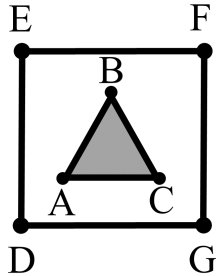
(b)



$$X = \{\sigma_1^1 + \sigma_1^0\}$$

$$= \{AB + C\}$$

(c)



$$X = \{\sigma_1^2$$

$$+ \sigma_1^1 + \sigma_2^1 + \sigma_3^1 + \sigma_4^1\}$$

$$= \{ABC$$

$$+ DE + EF + FG + GD\}$$

Vietoris-Rips Complex

Hypothetical sheets of rubber are ideal tools for reasoning about topological equivalency, but poor representations for computation. When data is used in a computer program it must be stored in a fixed size object (typically a matrix of real numbers), where a sheet of rubber gives too much ambiguity to be represented concisely. Instead, combinatorial objects must be used, allowing us to fix the geometry of a sheet of rubber and describe the topology of the resultant space. These objects are called simplicial complexes, which are triangulations of a representative pointset of a space.

There are multiple definitions for the construction of a simplicial complex, but this work focuses on the Vietoris-Rips (VR) complex²³. The VR-complex is a method of describing how the edges of a connected network should generate faces to form a geometric shape. When three points are path connected by edges and form a triangle, the face between these points is filled in. This is topologically equivalent to a deformed sheet, and therefore this can be continually deformed to a single point, and is thus topologically equivalent to a single connected component with $B_0 = 1$, $B_1 = 0$. Four or more points path connected in a cycle, are topologically equivalent to the sheet of rubber with a hole in it, where $B_0 = 1$, $B_1 = 1$, as these two spaces may be continuously deformed into one another. In these combinatorial spaces, paths of more than 3 nodes are said to form a cycle if they return to a starting node, and each occurrence of such cyclic features will increment B_1 .

A simplicial complex is a combinatorial object, as each complex may be built up from lower dimensional simplices. A 0-simplex is the 0-dimensional point. Connecting two 0-simplices by a line creates a 1-simplex, and connecting three 1-simplices by their vertices forms a 2-simplex, or a planar triangle. Note that the 2-simplex is formed by joining together 1-simplices by their 0-simplex faces. Similarly, the 3-simplex (the tetrahedron) is created by gluing four 2-simplices together by their 1-simplex faces, and k -dimensional analogues may be generated by gluing together $(k-1)$ -

Figure 5: Three Vietoris-Rips complexes and their combinatorial representations. The simply connected space in (a) gives a single 3-simplex consisting of each of the points. The disconnected 0-simplex and 1-simplex in (b) may be represented by the sum of these simplices. The disconnected 2-simplex and chain of 1-simplices in (c) is given by the summation of each individual simplex.

simplices by their $(k-2)$ -simplex faces. A set of $k+1$ points in \mathbb{R}^k fully connected by edges will form a k -simplex, denoted by σ^k . A finite collection of simplices which form the simplicial complex of a topological space is given as the set $X = \{\sigma_i^k\}$ where i is used to index each of the simplices in each dimension (Figure 5).

We may treat the simplices of a simplicial complex algebraically, where each k -simplex is represented by the union of the algebraic labels that make up that simplex. A simplicial complex is then represented as the sum of all the simplices in the complex. Whilst not a necessary consideration for the examples in Figure 5, it should be noted that in simplicial geometry, this will take place under modulo 2 arithmetic, which is to say:

$$A+A=2A=(2 \% 2)A=0A=0$$

$$A+A+A=3A=(3 \% 2)A=1A=A$$

By visual analogy, if we “add” a simplex to itself, this has the effect of cancelling both objects out.

The boundary operator, ∂_k , of a k -simplex, σ^k , takes σ^k and maps it to the $k-1$ dimensional faces of the simplices which make up the boundary of the convex hull. For example, the boundary of the 3-simplex, $\sigma^3 = ABCD$, would return the four 2-simplices which make up the faces of the tetrahedron.

$$\begin{aligned}\partial_3 \sigma^3 &= \partial_3(\{ABCD\}) \\ &= \{ABC + ABD + ACD + BCD\}\end{aligned}$$

The boundary ∂_2 applied to the space formed of two 2-simplices, $\sigma^2 = \{ABC + DEF\}$ gives the six 1-simplices which make up these two triangles.

$$\begin{aligned}\partial_2 \sigma^2 &= \partial_2(\{ABC + DEF\}) \\ &= \{AB + AC + BC + DE + DF + EF\}\end{aligned}$$

A property that should be noted is that applying two subsequent boundary operators of decrementing dimensionality ∂_k , and ∂_{k-1} to a k -simplex will always lead to an empty set. The reasoning

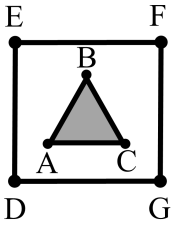
$$\begin{aligned}& \text{(a)} \\ & \partial_2 \left(\begin{array}{c} \text{B} \\ \diagup \quad \diagdown \\ \text{A} \quad \text{C} \end{array} \right) \\ &= \begin{array}{c} \text{B} \\ \diagup \\ \text{A} \end{array} + \begin{array}{c} \text{A} \quad \text{C} \end{array} \\ &+ \begin{array}{c} \text{B} \\ \diagdown \\ \text{C} \end{array}\end{aligned}$$

$$\begin{aligned}& \text{(b)} \\ & \partial_1 \left(\begin{array}{c} \text{B} \\ \diagup \\ \text{A} \end{array} + \begin{array}{c} \text{A} \quad \text{C} \end{array} + \begin{array}{c} \text{B} \\ \diagdown \\ \text{C} \end{array} \right) \\ &= \text{A} + \text{A} + \text{B} + \text{B} + \text{C} + \text{C} \\ &= \{\}\end{aligned}$$

Figure 6: In (a) the ∂_2 boundary operator is applied to the 2-simplex ABC to give the three 1-simplices which make up this boundary. By applying the ∂_1 boundary operator to these 1-simplices in (b), we see that each of the resulting components will cancel out under modulo 2 arithmetic.

behind this is simple, as the faces of any $(k-1)$ -simplex must be contained in exactly two k -faces of σ^k and thus cancel out under modulo 2 arithmetic. For example, applying ∂_2 followed by ∂_1 on the 2-simplex $\sigma^2 = \{ABC\}$ (Figure 6) gives:

$$\begin{aligned}\partial_1 \partial_2 \sigma^2 &= \partial_1(\{AB + AC + BC\}) \\ &= \{A + B + A + C + B + C\} \\ &= \{2A + 2B + 2C\} \\ &= \{\}\end{aligned}$$



$$X = \{ABC + DE + EF + FG + GD\}$$

$$\partial_2(X) = AB + AC + BC$$

$$\partial_2(X) = \begin{matrix} & ABC \\ \begin{matrix} AB \\ AC \\ BC \\ DE \\ EF \\ FG \\ GD \end{matrix} & \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \end{matrix}$$

Figure 7: The ∂_2 boundary operator may only operate on 2-simplices. As the VR-complex X has only one 2-simplex in there may be only one column in the corresponding matrix for this operation. There are seven 1-simplices in X and thus 7 rows in the ∂_2 boundary matrix. As ∂_2 takes ABC to $AB+AC+BC$ there is a 1 in these rows with zeros elsewhere.

Boundary Matrices

Mapping functions can be represented by matrices which represent the specific operations performed by a specified function. A boundary map tells us how a boundary operator ∂_k will take the k -simplices in a space X , σ_i^k , to their associated faces, σ_i^{k-1} . Let us demonstrate this process with a guided example before giving the more general mathematical definition.

The boundary map can take the form of a matrix with values in \mathbb{Z}^2 , i.e. $\{0, 1\}$. There is a column in the matrix for each of the k -simplices in the complex, with a row in the matrix for each of the $(k-1)$ -simplices in the complex. If a $(k-1)$ -simplex is the face of a k -simplex, then there will be a 1 in the corresponding entry of the matrix, with a 0 otherwise.

Applying this operation on the topological space $\{ABC+DE+EF+FG+GD\}$ (Figure 7) leads to a boundary matrix ∂_2 that can only operate on the 2-simplex ABC , taking it to $AB + AC + BC$. Therefore the corresponding boundary map will include a 1 in these rows, with 0 for the 1-simplices that make up the square, as these are not on the boundary of a 2-simplex. As there is only a single 2-simplex, the map is a column vector, with 7 rows, one for each 1-simplex in X .

Constructing the boundary matrix ∂_1 of X produces a map which takes each of these 1-simplices, or edges between points, to their boundary 0-simplices, or points. The corresponding map of ∂_1 for

X will thus have 7 columns, one for each of the edges in the space, and 7 rows, one for each of the points in the space.

$$\partial_1(X) = \begin{array}{c} \begin{matrix} & AB & AC & BC & DE & EF & FG & GD \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \\ F \\ G \end{matrix} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

The 0-dimensional points cannot be reduced any further. Applying the ∂_0 boundary map to the space will take each of these 0-simplices to the empty set, \emptyset , as there is nothing else they could possibly be mapped to. This will always take the form of a row vector with the number of columns equal to the number of points in the space:

$$\partial_0(X) = \emptyset \begin{bmatrix} A & B & C & D & E & F & G \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Computing Betti Numbers

Computational homology allows us to take a VR-complex and calculate the Betti numbers of the space in reasonable time. In the previously stated topological space X, we can visually attest that B_0 is 2 and B_1 is 1. This can be computed from the sequence of boundary maps $\partial_0, \partial_1, \partial_2$. We shall first refresh on some terminology.

In each of our boundary maps, the domain of ∂_k is the set of k -simplices that the map operates on. The image of a function is the set of all output values that a function may produce which are not

equal to 0. For the boundary map ∂_2 on X we know that $\{AB + AC + BC\}$ is in the image of the map, as it is non-zero and ABC is in the domain of ∂_2 .

The kernel of a function is the set of all inputs which get taken to the empty set \emptyset by the function. For the boundary operator ∂_1 operating on X , we know that $\partial_1(AB + AC + BC) = \{\}$, and thus $\{AB + AC + BC\}$ is in the kernel of ∂_1 . This can be attested by summing the corresponding column operations of the matrix boundary map of ∂_1 , noting the use of modulo 2 arithmetic. The resultant map is a 0-vector, which means that applying this to any input would lead to 0 (Figure 8).

$$\begin{array}{ccc}
 \text{AB} & \text{AC} & \text{AB} + \text{AC} \\
 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & + \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
 \end{array}$$

$$\begin{array}{ccc}
 \text{AB} + \text{AC} & \text{BC} & \text{AB} + \text{AC} + \text{BC} \\
 \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & + \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} & = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}
 \end{array}$$

Whenever ∂_k is applied to a valid complex it will have a set of $(k-1)$ -simplices in the image. In the corresponding ∂_{k-1} boundary matrix, at least one of the columns associated with the features in the summation of simplices will be reducible to the 0-vector, as $\partial_{k-1}\partial_k\sigma^k = \{\}$. This means that the output of ∂_k operator will always be in the kernel of the ∂_{k-1} operator on the same simplex.

Figure 8: We can sum the first and second columns of $\partial_1(X)$ using standard matrix operations. We see that by adding this column vector to BC the result is a zero-vector and thus $AB+AC+BC$ must be in the kernel of the transform.

There are however $(k-1)$ -simplices which do not form a face of a k -simplex that may also be in the kernel of a map, such as $\{DE+EF+FG+GD\}$ being in the kernel of $\partial_{k-1}(X)$. In fact any chain of $(k-1)$ -simplices will be in the kernel of $\partial_{k-1}(X)$. The sets that each of the boundary maps that operate on each of the k -simplices may be shown diagrammatically to exemplify this process on the space X , given in Figure 9.

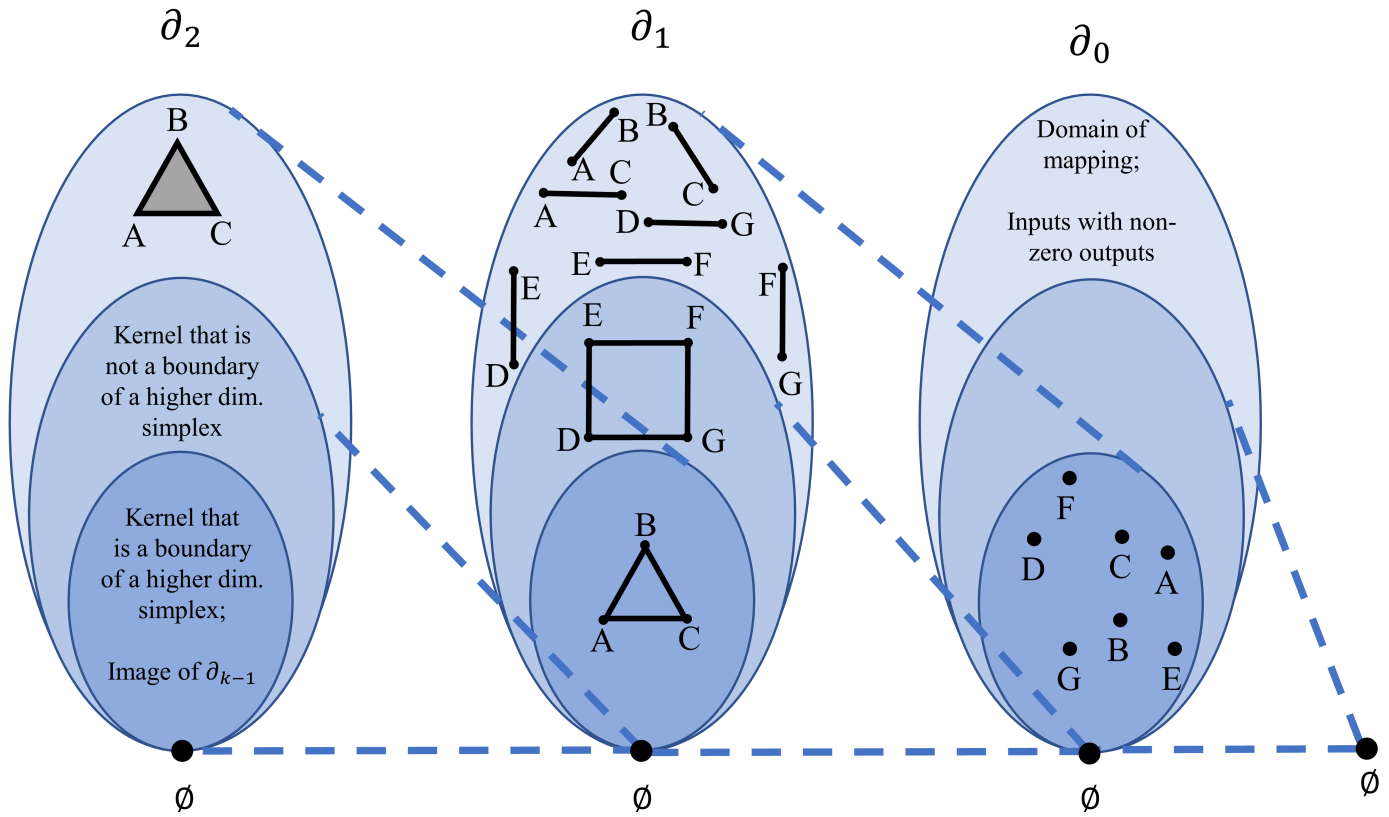


Figure 9: Each of the components of the space X and the result of applying each of the boundary operators. All cycles will be in the kernel, as the boundary will map this to zero.

We see that any chain of k -simplices will be in the kernel of ∂_k as the resulting reduction to the boundary faces will each cancel out. We also note that certain inputs in the image may lead to the same non-zero output, such as $\partial_1(AB) = A + B$, and $\partial_1(AC + BC) = A + B + 2C = A + B$. To find the true size (or rank) of the image

when all linear combinations of the inputs have been taken into account, we may examine the boundary mapping matrix, as the rank of the image of ∂_k will be equal to the rank of the matrix ∂_k .

The rank of a matrix is equal to the number of linearly independent columns. We previously saw how $AB + AC + BC$ is equal to the zero vector, and by a similar process we may show that $DE + EF + FG + GD$ is equal to the zero vector. Thus the boundary matrix for ∂_1 may in fact be simplified to a simpler basis via column Gaussian elimination.

$$\partial_1(X) = \begin{array}{c} \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \end{array} \begin{array}{c} \text{AB} \quad \text{AC} \quad \begin{array}{c} \text{AB+} \\ \text{AC+} \\ \text{BC} \end{array} \quad \text{DE} \quad \text{EF} \quad \text{FG} \quad \begin{array}{c} \text{GD+} \\ \text{DE+} \\ \text{EF+} \\ \text{FG} \end{array} \end{array} \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

As this matrix has 5 linearly independent column vectors, the rank of the image of ∂_1 is 5. For the map ∂_0 , we see that every input must be mapped to the empty set \emptyset , and thus the rank of the kernel is simply equal to the number of columns, which is 7. With these two numbers we may now introduce the core formula that will be used to derive Betti numbers:

$$B_k = \text{rank}(\text{kernel}(\partial_k)) - \text{rank}(\text{image}(\partial_{k+1}))$$

To find the number of connected components, which gives the first Betti number, B_0 , we take the rank of the kernel of ∂_0 and subtract the rank of the image of ∂_1 .

$$\begin{aligned}
B_0 &= \text{rank}(\text{kernel}(\partial_0)) - \text{rank}(\text{image}(\partial_1)) \\
&= 7 - 5 \\
&= 2.
\end{aligned}$$

This same formula and process may be continued to find the number of holes, B_1 . We know that the dimension of the kernel of ∂_1 is equal to the number of zero-vectors in the map, and as this matrix has two zero vector columns, the dimension of the kernel of ∂_1 is equal to 2. The boundary map ∂_2 is a non-zero column matrix, thus the dimension of this matrix must be 1. Therefore $B_1 = 2 - 1 = 1$, which is in agreement with observation.

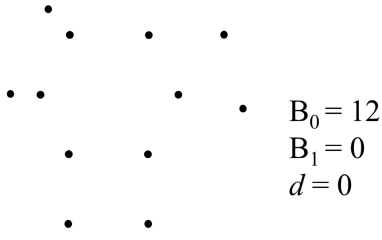
Although not presented here, this process may be continued into higher dimensional spaces with any number of points and topological features by simply calculating the boundary matrices and finding the dimensions of the associated images and kernels of these maps. When we compute Betti numbers in this work it should be noted that the only output of the algorithm is the two Betti numbers, B_0 and B_1 .

Persistent Homology of Finite Spaces

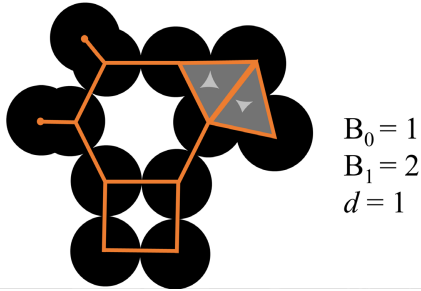
Betti numbers are clearly crude distinguishers. Whilst the topological information of a VR-complex may be used to describe pointsets, it does not provide much information. Two highly dissimilar pointsets could have identical Betti numbers with the appropriate choice of edges. It is therefore important to make sure that the chosen edge set is representative of the pointset that we are describing. This introduces a second problem with using Betti numbers as structural descriptors for ionic systems. In molecular chemistry, known covalent bonds provide a natural method of ascribing edges between atomic nodes, but what is the corresponding edge set of an ionic crystalline solid?

Approaches such as CGCNN will typically add all edges shared with neighbouring atoms within a neighbourhood radius of 8\AA . Whilst this will capture the local physical connections, and has demonstrated great success at predicting chemical properties, it

(a)



(b)



(c)

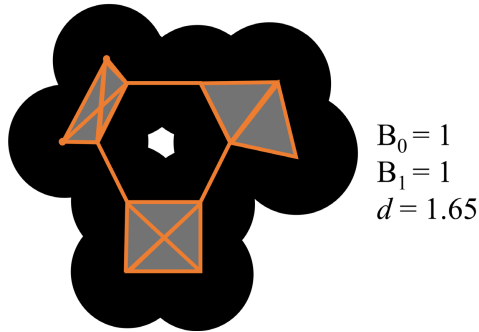


Figure 10: A 0-dimensional pointset has B_0 equal to the number of points in the set (a). As covering disks expand from these initial positions they will overlap and the topology of the space changes. At a certain diameter, all components will be connected (b). As the disks continue to grow (c), all topological components will eventually die leaving a single connected component.

discards the long range structure, giving rise to potential ambiguities. Here we describe a human-interpretable representation that reflects variations in the crystal structure, which has an associated metric between representations. Persistent homology is applied and expanded upon to further this goal. This examines how the topology of an edge set changes as the edge cut-off distance of a pointset is increased from $[0-\infty)$.

This may be visualised by expanding each of the points in the pointset from a 0-dimensional object into a solid sphere. As the diameter of the sphere grows the boundaries of neighbouring sphere will overlap. Each time two spheres overlap, these are topologically equivalent to the 1-dimensional line between the two nodes, and a 1-simplex is placed between the corresponding vertices in an associated VR-complex.

This gives a sequence of nested simplicial complexes starting from the pointset, with edges added until the fully connected graph is reached. Each generated complex is a subspace of the next complex in the sequence, and we may say that one is mapped to the next via an inclusion map. This sequence of nested simplicial complexes is referred to as a filtration of the final VR-complex (Figure 10).

After the points are fully connected by covering spheres, increasing the diameter further will not cause any extra edges to be added to the complex, and the process may terminate. Computing the homology of each VR-complex in the filtration returns a series of Betti numbers, which naturally provides more information than taking the Betti numbers at a fixed radius. Combining this knowledge with the associated radius each Betti number increases or decreases, allows us to recapture some of the original geometric information of the space.

Each time two points connect we say that a feature in B_0 has died, as topologically there is one less connected component in the space. Each filtration necessarily starts with $B_0 = n_points$. The features die one by one as each point merges with the wider

structure, until a fully connected complex with $B_0 = 1$. When a set of spheres overlap to form a space equivalent to S^1 , a topological feature is added to H_1 , as a space is formed that cannot be continuously mapped to a single connected component. As the diameter of the covering disks continue to grow, cycles will be “filled in” as more points form connected triangulations with one another. Eventually, the VR-complex triangulates the space fully, and all associated topological features in H_0 and H_1 will die barring a single H_0 feature, representing the fully connected space, with $B_0=1$.

When this process is complete, the birth and death times of Betti numbers can be tracked. It is these birth and death times, or persistence intervals, that shall be the main focus of interest. Persistence intervals track the birth and death times of each of the topological features in the space. Each time a Betti number increments, persistence intervals with the associated birth times are added. Each time a Betti number decrements a persistence interval with no currently assigned death time is chosen to die. This is chosen via the elder rule ²⁴, which selects the interval with the largest available birth time to die, which ensures that features born at the start of the filtration persist until they are completely “filled in”. Features with the largest persistence (*death – birth*) can be said to be the most topologically significant. Shorter lifetime features describe smaller cycles that can be found in the space. The most salient topological features are thus extracted from pointcloud data, which may be used for direct comparison using metrics, or used to represent the topology of the pointcloud on a diagrammatic plot, which gives a visual familiarity to spaces that are notoriously difficult to visualise.

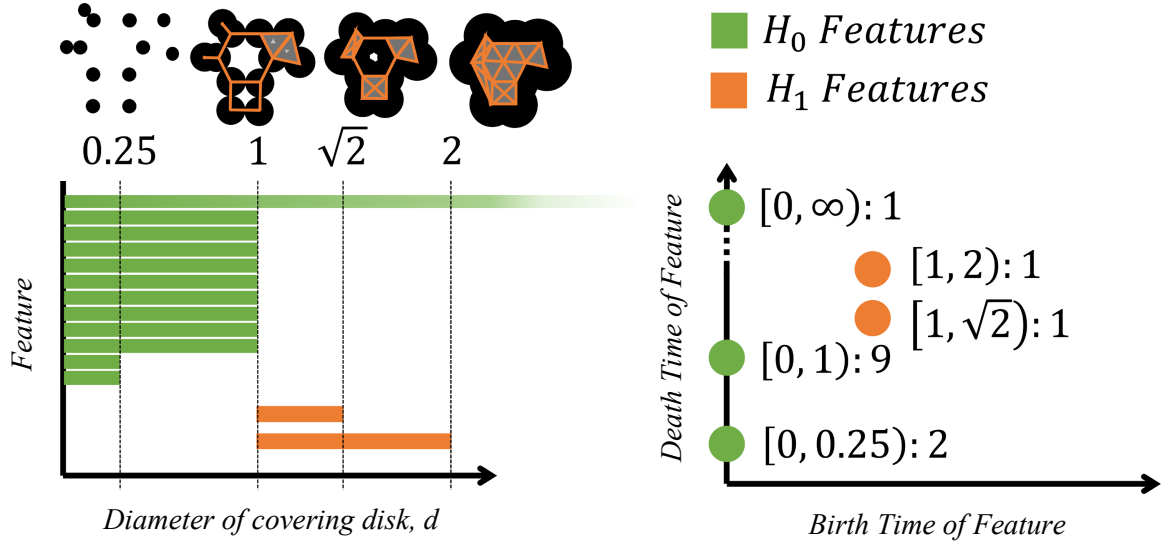


Figure 11: A persistence barcode (left) and persistence diagram (right) of the same space. On the persistence barcode each topological feature is described by a bar, with the birth and death times of the feature marked by the bars features. On the persistence diagram, each feature is represented by a point, with the birth and death times on separate axes. The co-ordinates and multiplicity of each point have been overlaid.

Persistence intervals are commonly plotted on either a persistence barcode or a persistence diagram (Figure 11). The persistence barcode marks the cut-off radius on the x -axis over a filtration; for each connected component born at $d = 0$ (*i.e.* each point) a bar is added to the H_0 barcode. As d grows large enough to connect a point to its nearest neighbour, the corresponding bar for the point will die, as the space has one less feature. Each bar in H_0 dies with the corresponding death time of an interval, with the last single feature stretching to infinity, as increasing d further will not change the resultant topology. While all features in H_0 are born at $d = 0$, this cannot be the case for the higher dimensional features. Each cycle that is formed across the expansion of d has its birth and death times marked on a separate H_1 barcode. Both are typically given on the same plot, where the barcodes of pointclouds with few points can often be distinguished from one another by eye. For large pointclouds, barcodes are often difficult to interpret due to the many similar bars on one plot.

Persistence intervals may instead be represented by 2D scatter plots, called persistence diagrams (PDs). These contain one point for each bar on the bar code, with the x co-ordinate taken as the bars birth time, and the y co-ordinate taken as the death time. Each of the H_0 points fall on the y axes as these are all born at $d = 0$. The H_1 features are marked separately as points in the region bounded by $y > 0$ and $y > x$, as features cannot die before they are born. For this reason the persistence diagram is often accompanied by the line $x = y$. Points falling very close to this line either originate from topological noise introduced through errors in floating point geometry, or from very short lived topological features. The topological features with longer lifetimes will fall further from the diagonal, as these are born at a small value of x and survive until a high value of y . These long lifetime features are often observed to be the most representative of the global topology of a space.

The persistence diagram gives a concise way of visualising unstructured pointcloud data in a 2D plot, where shared structural motifs between pointclouds are often reflected by similar patterns in PDs. There are many metrics to compare two PDs, and in this work we will focus on applications of optimal transport (the earth movers distance). Persistent homology has previously shown itself to be a suitable technique in many atomic domains^{25,26}. Here we present some further modifications to the core theory which allows for a more stable PH representation for periodic pointclouds.

The first modification to the PD is merely one of convenience. As d is increased from $[0-\infty)$, at a critical point the simplicial complex will be fully connected and further increasing d to ∞ will not change the topology of the complex. This means that every generated set of persistence intervals contains the feature $[0-\infty)$ in the H_0 set. This introduces the usual complexities when dealing with ∞ , and is a shared characteristic across all PDs. This feature can thus be altered such that once the PD is calculated, the largest

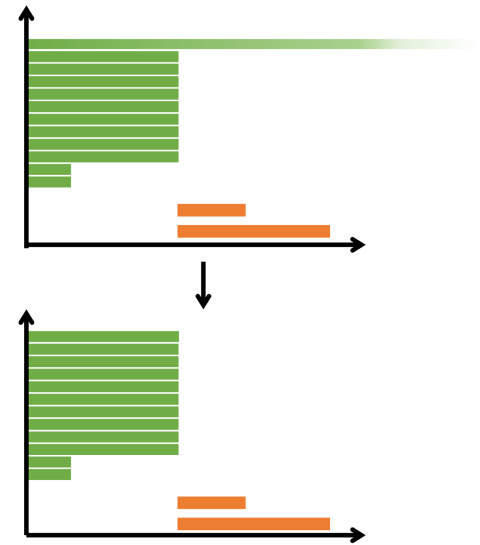


Figure 12: The persistence interval in H_0 with the birth-death time of $[0-\infty)$ has its death time constrained to the highest finite death time in the set.

death time (∞) is made equal to the second largest living feature in H_0 , giving a fully finite PD (Figure 12).

The second modification is built from a simple observation of periodic systems formed from lattices. Examining the PDs of the primitive lattices, it is apparent that the values of the intervals for each topological feature identified are independent of the choice of unit cell. Larger unit cell expansions will cause a greater number of features to appear on the barcode, but these will all start and end at the same time. From this observation we apply one of our main modifications to the theory to make it more suitable for periodic systems.

As opposed to keeping an entry for each topological feature that is observed, a *key: value* data structure is used to represent each set of the repeating features. Each of the unique $[birth, death)$ intervals in a set of features is extracted and used as a key, with the fractional ratio of the count of each interval in the set stored as the value (Figure 13). Applying this to crystal systems brings us closer to our desired goal of stability in the representation, as any integer expansion of the basis lattice for a simple system will lead to an identical PD. This representation will be referred to as the fractional persistence diagram, or FPD.

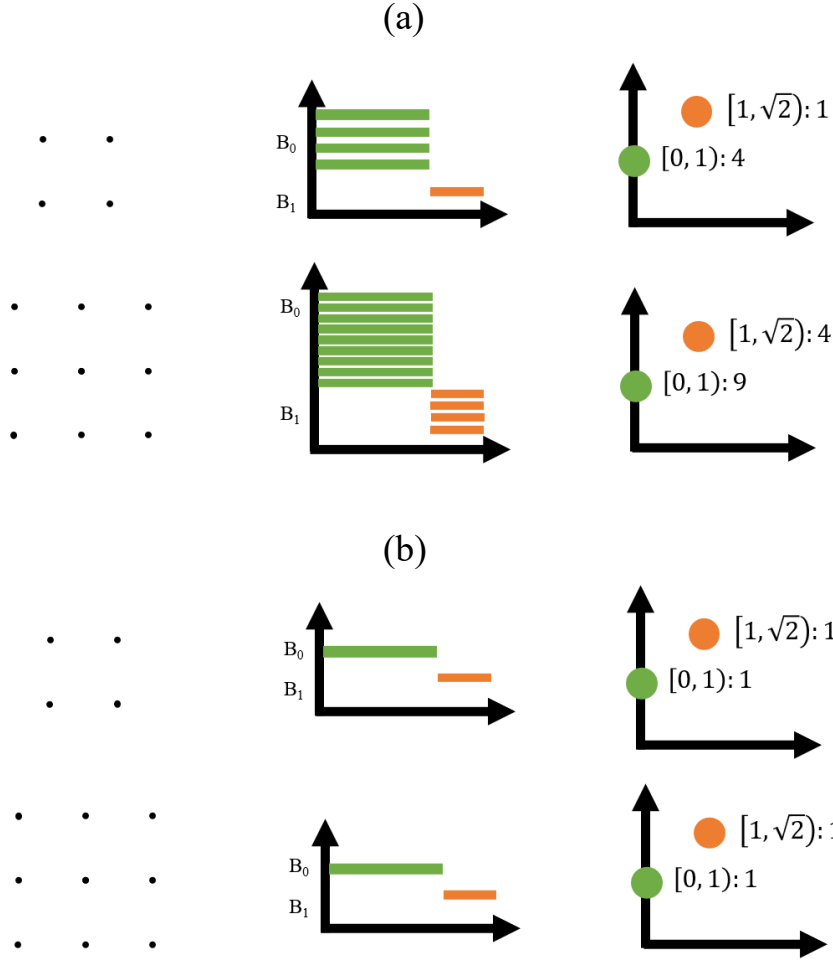


Figure 13: The square lattice in (a) gives the same unique persistence intervals as the expanded cell, with the multiplicity of these features also increasing. In this example, taking the fractional ratio of the unique persistence intervals across each homology set (b) gives a fractional PD which does not vary with unit cell expansions.

FPDs may be generated for the primitive Bravais lattices in 2-dimensions, and these representations can be arranged into a classification which reflects observable changes in the periodic pointcloud. Being simple periodic spaces, these are natural objects to explore how structural deformations are reflected by deviations in FPD. It may be noted that features in each representation vary proportionally, with the lattice vectors, and are fully parameterised by these values.

Primitive Bravais Lattice Classification

First consider the lattice for the simplest periodic pattern, repeating points in a straight line, which can be described by a lattice where $a = b$, and $\alpha = 0$. Expanding covering disks will give one topological feature on the diagram in B_0 for each point (which will collapse to a single feature of ratio 1), before these merge at

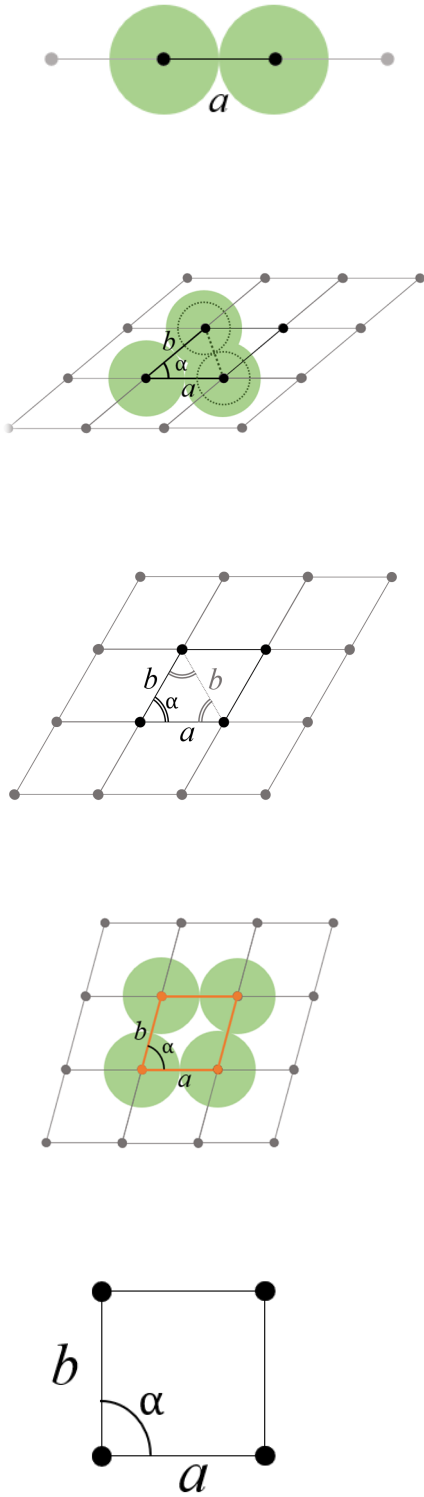


Figure 14: For the 2-dimensional lattice with equal vector lengths ($a=b$) this will have a single component which dies at diameter a when $\alpha = 0$. Increasing this angle will lead to a second component appearing which will increase in death time until the space forms a tiling equilateral triangle with a single feature in the FPD. Further increasing α gives a cycle in the space which will increase in death time until the square lattice is formed.

diameter a . This space clearly has no cycles which form, and thus the B_1 set is empty. Increasing the lattice angle α , allows an oblique lattice to be generated, where this time covering disks will cause components to connect at diameter $2a \sin\left(\frac{\alpha}{2}\right)$, before merging at diameter a . This causes a second interval to expand from 0 in B_0 , but as the space forms a complete triangulation at diameter a , no cycles form.

The death time of this second feature increases proportionally with α until the point when $2a \sin\left(\frac{\alpha}{2}\right) = a$, as we have reached a hexagonal lattice. In the primitive hexagonal lattice, all topological features merge at diameter a , and the FPD representation is thus indistinguishable to the linear FPD.

Further increasing α leads to a second oblique cell, but with a distinguishable FPD from the previous set of oblique lattices. In this lattice, all components will connect at diameter a , but this will not form a triangulation of the space. Instead, periodic cycles are formed, which die when the diameter of the covering disks is equal to the length of the longest diagonal of the unit-cell, $2a \sin\left(\frac{\alpha}{2}\right)$.

Increasing α to the angle $\frac{\pi}{2}$ eventually gives a square lattice, and as this forms a right angle triangle with the diagonals of the unit-cell, the topological feature associated with the cycle's death time will increase up to $\sqrt{a^2}$ (Figure 14). The process of increasing the lattice angle may be continued from a square cell, where the same sequence of oblique and hexagonal pointclouds and their associated FPDs are generated, in reverse order, arising from the observed symmetry that equivalent lattices may be reduced to one another.

Alternately, a separate lattice transformation can be applied, increasing the length of one of the lattice vectors, where it may be assumed without loss of generality that $a < b$. Applying this to the square lattice leads to an rectangular lattice which causes a second feature to split off from the first feature in H_0 which will die at

diameter b , with the feature in H_1 having its death time increase to $\sqrt{a^2 + b^2}$ (Figure 15). Increasing the length of b will lead to other orthorhombic cells, and thus only affect the lifetimes of the topological features in each dimension, not their count or their arrangement in regards to each features overlap in the resultant FPD.

Varying α from an rectangular cell will lead to more oblique cells, with FPDs that have similarities to those that neighbour the square lattice, but bearing an extra feature in H_0 which dies at diameter b . The lifetime of the cycle in H_1 will diminish as α moves away from $\frac{\pi}{2}$ until α reaches the isosceles angle (*iso*) such that the periodic pointcloud resembles a tiling of tessellating isosceles triangles. Here, two features form in H_0 which will die at diameters a and b , but as the leading diagonal of the unit cell is equal to the edge length b , no cycles form as the space is a complete triangulation (Figure 16). Further decreasing the lattice angle will lead to a cycle forming in H_1 , but this time instead of forming at diameter b , this will be created when covering disks overlap across the cell diagonal, and die at diameter b .

Even when $b \gg a$, decreasing α will eventually lead to a pointcloud which could be reduced to a rectangular or square lattice, which possesses an identical FPD to the pointcloud generated by the unreduced lattice. This occurs when α reaches the critical point, *iso*’, where the triangle constructed by the lattice vectors and the opposing diagonal is once again isosceles. This process will cycle through each of the equivalent lattices potentially many times if $b \gg a$, until a lattice which is equivalent to the hexagonal lattice is generated. At this point, there are no cycles in the FPD as this is a complete triangulation of the space, and further decreasing the lattice angle will simply lower the death time of the second topological feature until it approaches 0 and we have returned to a linear representation. This sequence of operations may be shown diagrammatically, as given in Figure 17.

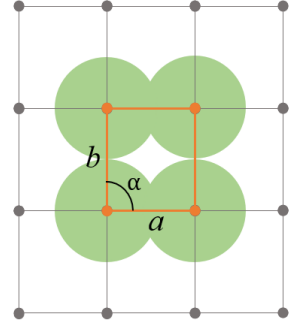


Figure 15: The orthorhombic cell forms a cycle at diameter b .

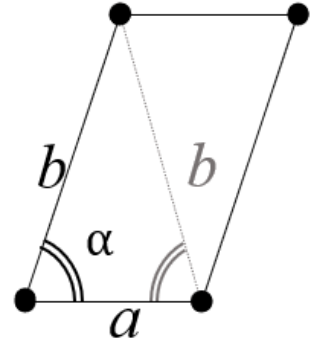


Figure 16: Decreasing the lattice angle α will eventually lead to an isosceles triangle which does not form cycles.

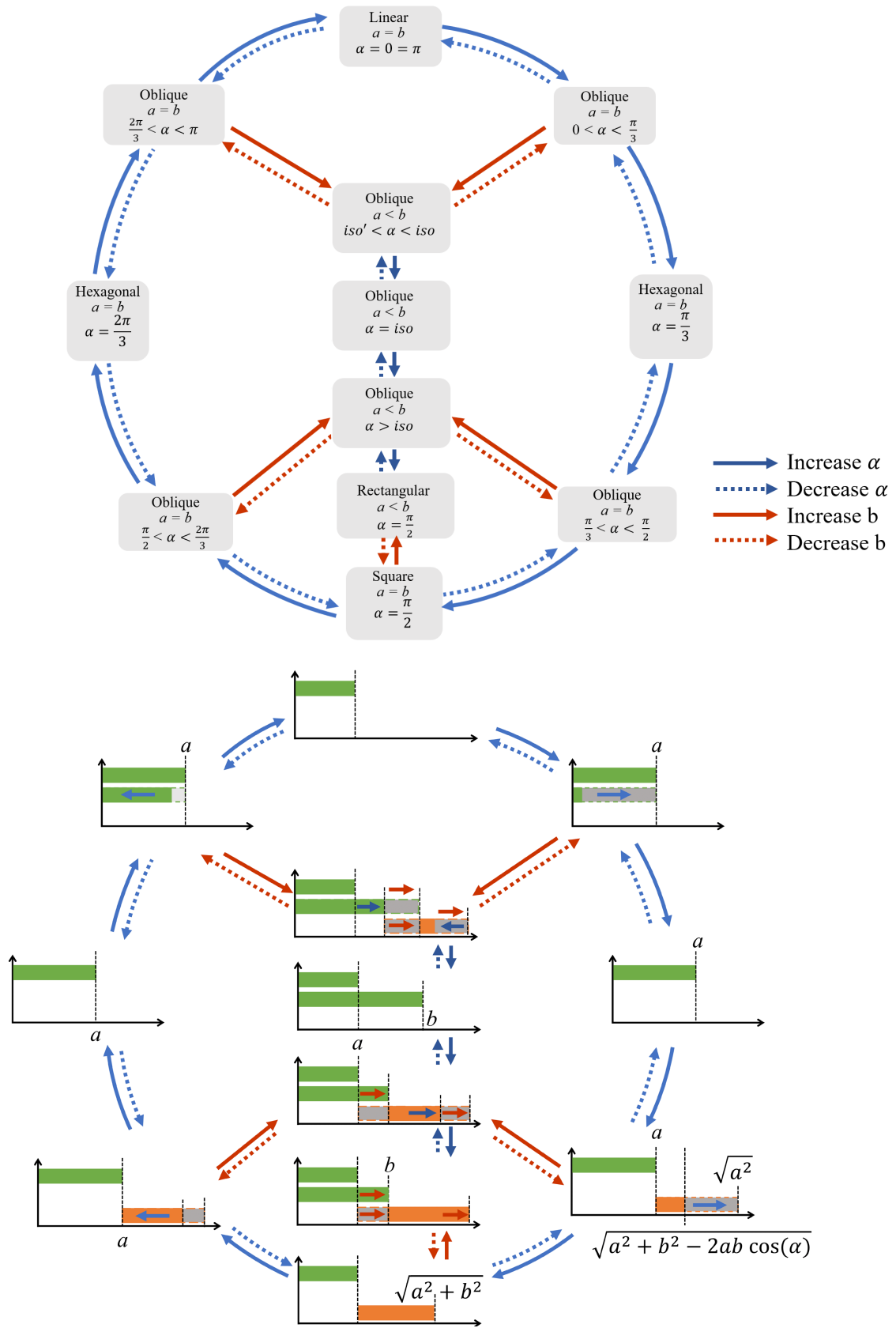


Figure 17: The primitive Bravais lattices can be defined by their lattice vectors and lattice angle, and arranged into a series of lattice transforms (upper). Each lattice transform will lead to a corresponding change in the birth or death times of a feature in the resulting FPD, as indicated by the arrows overlaid on each diagram. Inverse (dashed) operations may be followed by reversing the direction of the arrows on each FPD.

By increasing a lattice vector or angle, the resultant FPD has a corresponding deviation to the original FPD. This is typically reflected by the bar boundaries increasing or decreasing and the masses of bars merging or splitting in two. To create Figure 17, an FPD was constructed for each of the primitive lattices, then manually arranged to ensure continuity, with the corresponding lattice transformations derived *a posteriori*.

For the 2-dimensional Bravais lattices we consider the hexagonal lattice ($a=b$, $\alpha=60^\circ$) as the lattice with the simplest topology. The repeating pattern of equilateral triangles ensures that no cycles can form in the persistence diagram. The connected components of the pointcloud join at $d = a$, and the resultant sheet is fully connected, leaving a single point in H_0 and no points in H_1 . Increasing the lattice angle, α , or increasing the magnitude of either of the lattice vectors leads to monoclinic lattices.

Consistently arriving at a reduced cell lattice representation of a periodic pointcloud for all feasible periodic systems is a solved problem^{27,28}, but generating reduced lattice vectors in a stable fashion is not. The outputs of lattice reduction algorithms may change considerably under small variations in structural motifs, and care must be taken to ensure a structural descriptor does not return noticeably different descriptors for the same underlying pointcloud. Until now we have considered finite or simple systems where these details could be disregarded. Unfortunately, the periodic boundary conditions imposed by the structures of solid state chemistry introduce further complexities which break invariance when using the necessarily finite representations reported by experimentalists in crystallographic information files (cifs).

Persistent Homology Shortfalls for Periodic Pointsets

For the primitive Bravais lattices, it was assumed that any choice of unit cell would generate an identical FPD. There is however, a simple counter example which will return different FPDs for different choices of input lattice vectors, being the last of the planar Bravais groups, the face centred rectangular lattice (Figure 18).

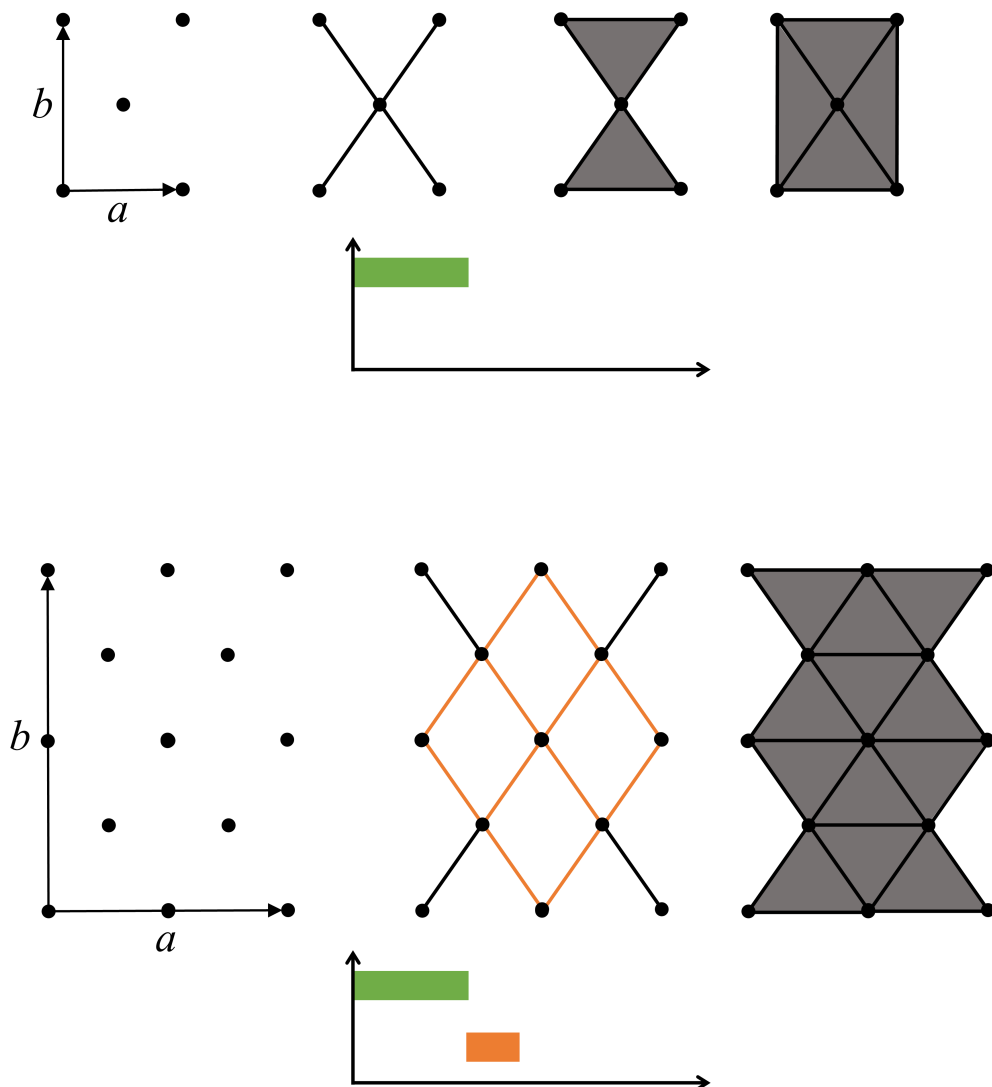


Figure 18: The associated filtration and FPD of the 1-cell (upper) and the 2-cell (lower) expansions of the centred rectangular lattice. Unlike the other primitive Bravais lattices, expanding the lattice vectors of the face centred rectangular lattice can lead to FPDs which have different persistence intervals in the resulting FPD.

The filtration of the pointcloud generated by the 1-cell fails to capture the longer range topology that is contained in the periodic structure. An expansion to the 2-cell reveals the cycles present in the larger pointcloud, demonstrating this lattice and its associated FPD are equivalent to the primitive monoclinic lattice. A naïve proposal to resolve this issue could be to simply use a larger expansion, such as the 3-cell, to ensure that all internal topology is captured. Unfortunately, for more complex atomic motifs with greater numbers of points contained in the unit cell, different choices of unit cell vectors have the potential to generate differing FPDs.

As an example, the circular motif in the centre of a tetragonal lattice returns a fractional PD which correctly captures the two repeating cycles when the lattice does not break up the circular pattern. If we translate the lattice across the same pointcloud we may generate the same space, but the new disjoint motif which crosses the lattice vector causes an extra topological feature to appear on the FPD, which is not present in the wider periodic structure, shown in Figure 19.

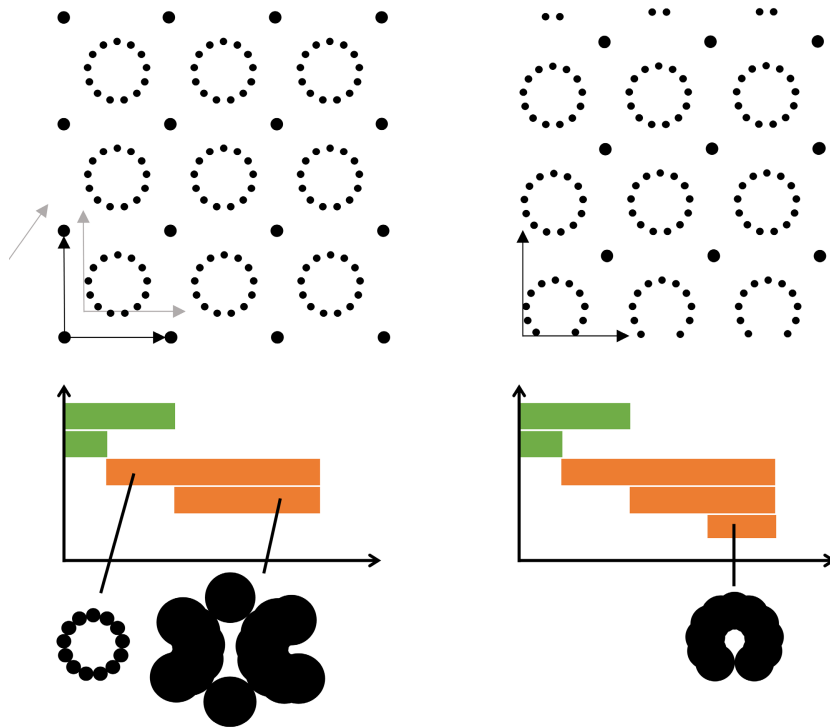


Figure 19: The repeating circular pattern on the left will have an associated FPD, but sliding the lattice vectors across the space (shown in grey) until they intersect a motif can cause fragmented features to appear in the diagram that aren't present in the repeating structure (right).

We shall refer to these additional features as fragmented boundary features, artifacts of the process of creating finite realisations of periodic structures. It can be shown from the axioms of metric spaces that the distance between an object to itself must be 0. Here, the same underlying pointcloud has created two representations which must have a non-negligible distance greater than 0 when measured under almost any metric. This representation needs further development to achieve consistency.

Percifter

Percifter is a process to generate stable topological descriptors from cifs, with a focus on removing the topological noise generated from boundary features. Luckily, there is a clear solution to removing boundary features. Simply take the unit cell expansion up to ∞ , compute the FPD of the space, and the ratios of the boundary features will be equal to 0. Unfortunately, this is not simple on a computer with finite memory.

Even for lattice expansions up to the 10-cell, it is not feasible to compute the persistence diagram of a complex material in reasonable time with present technology. In fact, even 5-cell expansions remain impractical for high-throughput purposes, with complex motifs exceeding the typical memory limits of a modern workstation. Smaller unit cell expansions may be computed in a timely manner, but boundary features will make up a larger proportion of the final ratios in the representation. These features could simply be recognised as falling within an acceptable margin of error of the representation, but we propose the following technique to remove them.

Percifter builds upon the observation that increasing the expansion of the unit cell leads to a corresponding decrease in ratio for each of the boundary features. Marking the ratios of the features of FPDs across a sequence of three unit cell expansions allows us to take note of whether each topological feature is growing in overall proportion, or decreasing in proportion to the remaining topological features. For 3-dimensional volumetric space, periodic

features contained within the interior of the convex hull will be increasing at a rate of n^3 , whereas features found on the boundary of the surface will be increasing at a rate of at most $6n^2$. This is because there can be at most 6 repetitions of the boundary feature on the tiling polyhedra, and these ratios follow an n^2 increase.

To make this process work, we assume that the ratio of each of these features follows a function, which may be approximated by fitting a curve. By following the sequence of features across increasing expansion factors, x , we can fit the unknown variables a , b and c of the function, $f(x) = \pm \frac{a}{x-b} + c$ to predict what the ratio of a feature could be at expansion, x , such that $f(x) \approx \text{ratio}^x$. The limit of this fitted function is known to converge, and this limit is taken as x goes to ∞ , which is used as the final ratio of the feature in the representation.

This allows us to determine empirically which features are likely to be fragmented boundary features which are not present in the wider structure. Features with final ratios which fall below a predefined tolerance (less than 0.001 in this work) are removed from the diagram. The remaining features are scaled to ensure the total mass of the feature set is equal to 1. Given the potential confusion that could arise in notation when indexing each expansion factor by 1, 2, 3, or, a , b , c , or x , y , z , or α , β , γ , we refer to the index of each expansion in the series using the ‘I’ symbol, [I, II, III].

For each PD across the three expansions, we construct a separate key: value lookup. As with the FPD, the key is the (birth, death) time of each topological feature, rounded to $3d.p.$, and the value is the ratio of that feature. To simplify notation this is given as the ratio of the i -th topological feature in a set for a given expansion x , ratio_i^x . Thus, we may join the topological features across the expansions, marking whether the ratios of a topological feature are increasing or decreasing. Each feature will have a corresponding function f_i , that fits the output to each $\text{ratio}_i^{\text{I}}$, $\text{ratio}_i^{\text{II}}$, and $\text{ratio}_i^{\text{III}}$ from the associated expansion values. Once we have found the

optimal parameters for f_i which minimise the total error for $f_i(x) = ratio_i^x$, the limit of the i -th topological feature at an infinite expansion factor, $f_i(\infty)$, is taken as the final ratio of that topological feature. It is observed that constant, $\frac{1}{x}$, and $-\frac{1}{x}$ functions fit these series reasonably well, which take the general form:

$$f_i(x) = \begin{cases} \frac{a}{x-b} + c & , \text{ } ratio_i^I < ratio_i^{II} \\ ratio_i^I & , \text{ } ratio_i^I = ratio_i^{II} = ratio_i^{III} \\ -\frac{a}{x-b} + c & , \text{ } ratio_i^I > ratio_i^{II} \end{cases}$$

Where a , b , and c are parameters that are found through non-linear least squares fitting. This is an iterative process where random values are initially chosen for each parameter and refined at each step using the gradient between the observed and calculated outputs of the function to update each parameter, until each of these has converged to a stable value and altering these does not reduce the total error any further.

If the ratio of a topological feature is increasing ($ratio_i^I < ratio_i^{II}$) this must mean that it is an internal periodic feature. Similarly, if the ratio remains constant, this may also be assumed to be a periodic feature contained in the larger periodic structure. Conversely, if the i -th feature is decreasing in fractional makeup ($ratio_i^I > ratio_i^{II}$), this means that it is more likely to be a boundary feature. Periodic topological features found in the infinite pointcloud may simply be in different fractional makeup between the first and last expansion. Others however will be the remnants of fragmented motifs, which are not found in the periodic structure. These unwanted topological features should not be contained in the final representation.

As we follow the series of a boundary feature f_i , we see that $ratio_i^x$ tends towards, or falls below, 0 as x approaches ∞ . All such

features with a ratio below 0 (or sufficiently close to 0) are regarded as topological noise and removed from the PD. Each of the non-zero topological features in each set are used as keys in the final representation, and their ratios at infinity taken as the associated values, rescaled to give a total mass of 1. This final representation of a crystal structure will be referred to as the convergent persistence diagram, or CPD (Figure 20).

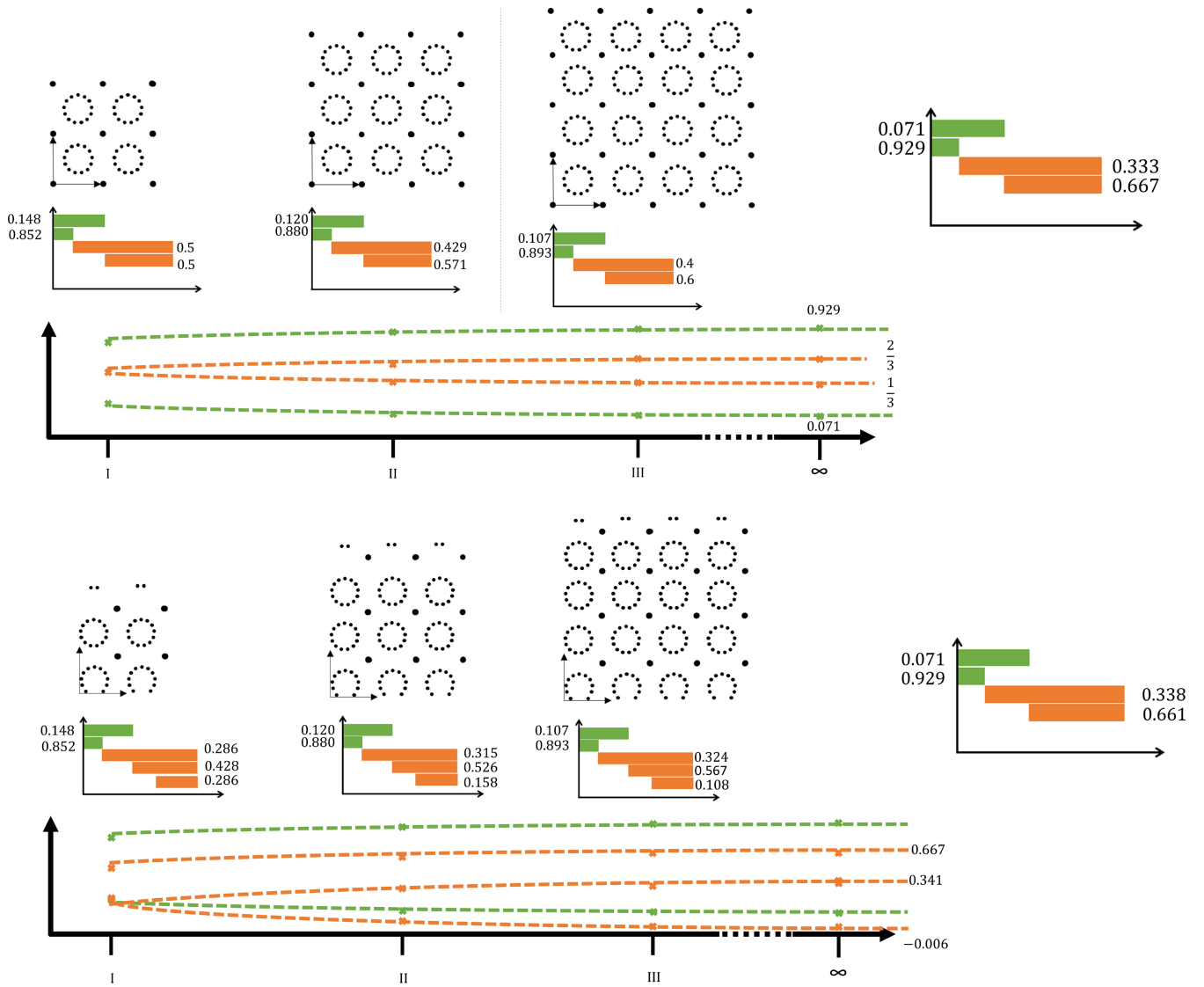


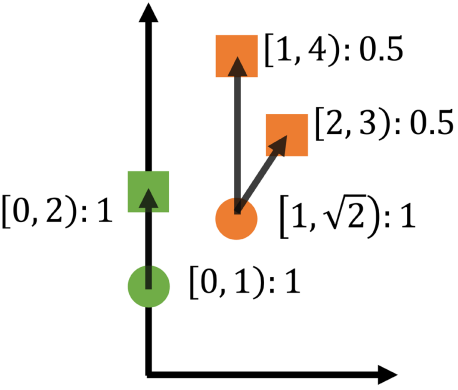
Figure 20: Using the same motif and lattice as Figure 19, we can see how scaling these unit cells to the 2, 3, and 4-cell expansions gives different ratios of each persistence interval. In the upper plot, four functions (one for each persistence interval) with the general formula $\frac{a}{x-b} + c$ have their a , b , and c parameters altered until a good approximation of the intervals fractional ratio as a function of cell expansion is obtained. The limit of the function at ∞ is taken as the convergent ratio of the feature. In the lower plot, we see that the function associated with the fragmented boundary feature has a negative value at ∞ , and thus may be removed from the final CPD.

The CPD gives a concise representation of the crystal structure, which can be said to be an interpretable descriptor from the ability to construct plots of CPDs as visual aids. It may be noted that in the previous example the two CPDs are not identical to one another. By definition this means that this implementation is not an invariant representation of the periodic pointset. Proving the convergence and stability of the descriptor remains an open question, although larger expansion values for I, II, and III, should lead to more accurate final ratios taken at ∞ . Resultant CPDs for lattices representing the same underlying pointcloud are typically measured as having a negligible distance from one another under standard metrics.

An established metric of similarity between persistence diagrams is one that is familiar to us: the earth movers distance (EMD), or the Wasserstein metric, as it is typically referred to in topological data analysis. Yet again, the analogy of computing the optimal method of transporting earth from one distribution into another remains valid when considering the distance between CPDs. Each of the points in each homology group of the CPD is considered as a weighted Dirac mass, the weight being the final ratio of the feature, with the total sum of masses in each set being equal to 1. The planar Euclidean distance between points on the CPD is taken as the distance between features, and the optimal transportation plan to assign mass is calculated.

This minimal quantity of work required to transform one distribution into another is taken as the distance between the set of intervals in H_0 and H_1 , giving two separate distances. Whilst each homology group should be considered independently, as connected components and cycles are topologically distinct objects, a metric should return a single real valued number. To ensure this metric can be used in downstream tasks, the two distances are averaged together, and returned as a final measure of structural similarity (Figure 21).

(a)



(b)

$$d(\bullet, \blacksquare)$$

$$= 1 \times \sqrt{(0-0)^2 + (2-1)^2}$$

$$= 1$$

$$d(\bullet, \blacksquare)$$

$$= 0.5 \times \sqrt{(1-1)^2 + (\sqrt{2}-4)^2}$$

$$+ 0.5 \times \sqrt{(1-2)^2 + (\sqrt{2}-3)^2}$$

$$= 2.052$$

(c)

$$d(\bullet, \blacksquare)$$

$$= \frac{1 + 2.052}{2}$$

$$= 1.526$$

Figure 21: Two CPDs are overlaid on one another (a), the first plotted by circles the second by squares. The Wasserstein distance is the cost to transport mass between points, and between these two diagrams must be considered for each homology group in turn (b). Once each of the Wasserstein distances have been computed, these may be averaged to give a total distance between the two diagrams (c).

Results and Discussion

Empirical Construction Time

Creating a persistence diagram is a computationally intensive process, on the order of $O(n^3)$, and the Percifter algorithm with a [2, 3, 4] expansion series must construct a 4-cell during execution, multiplying the number of input points contained in a cif file by 64. Despite the large inputs and time intensive procedures, the average running times are manageable for structures with fewer numbers of points, which is empirically shown by Figure 22. Whilst the average running time is lower than the theoretical worst case, at present it remains computationally costly to process structures in a high throughput manner with this expansion factor. Processing 137,438 structures each containing fewer than 176 atoms in the unit cell required 4,035 CPU hours in total.

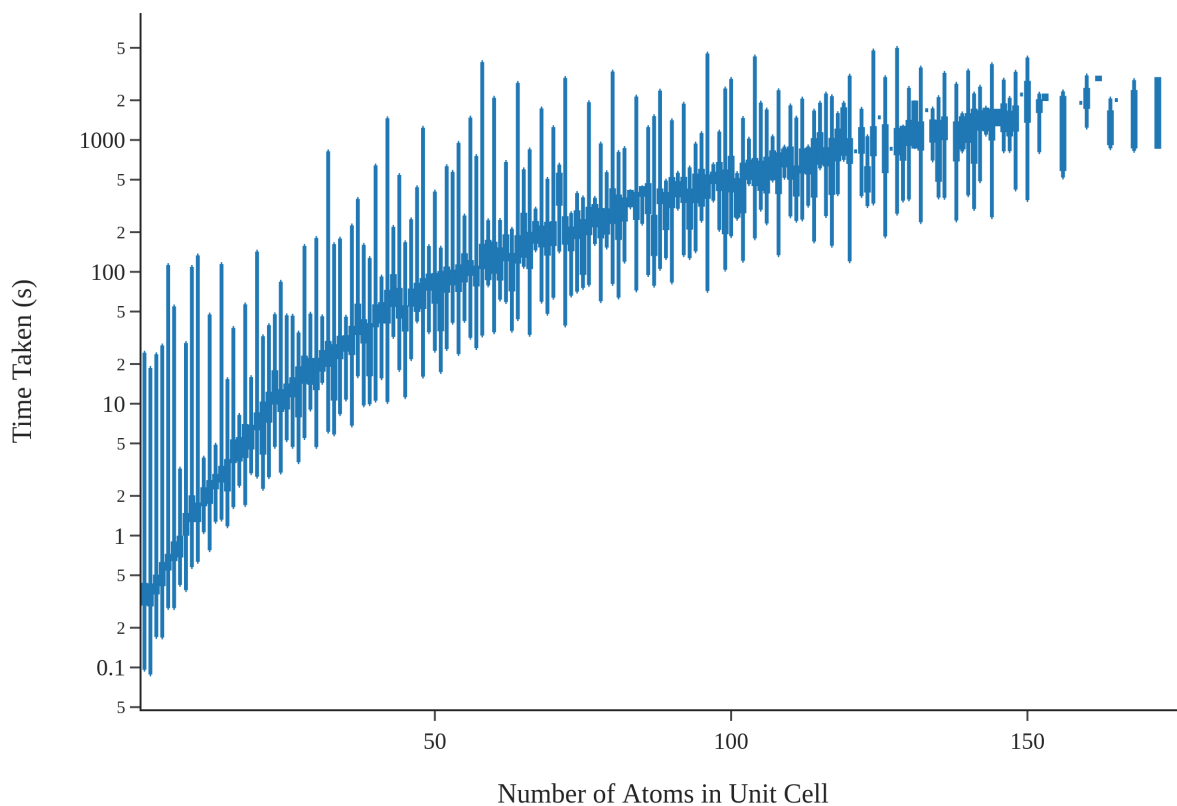


Figure 22: Empirical Construction Times to create the Percifter representations of 137,438 crystal structures using a [2, 3, 4] expansion.

This is an acceptably small running time for general applications, however high throughput analysis of larger cells remains difficult. Materials which could attain the most value from tuning the structural properties, such as porous materials and ionic conductors, typically contain many atoms in the unit cell due to their complex structures, necessitating the use of smaller expansion factors, such as [1, 2, 3]. The ripser algorithm²⁹ used to generate each persistent diagram underlying Percifter carries forward many matrix reduction operations. Although not implemented in the original library, this may be parallelised and run on GPUs³⁰ for further speedup, a possible avenue of future study for larger systems.

The sequence of unit cell expansions is a core hyperparameter of the Percifter algorithm. Longer sequences of larger expansion

factors will lead to smaller error in the final ratios. Unfortunately, the computational resources required to compute these diagrams for complex materials of interest is an ever present limiting factor. For these purposes we have explored different series of expansion factors in this work. For the simpler binary systems 2-cell, 3-cell, and 4-cell series ([2, 3, 4]) were initially used.

It is not possible to compute the 4-cell of a complex disordered electrolyte or a metal oxide framework in reasonable time. Therefore the [1, 2, 3] cell expansions were explored, with the restriction that all lattice angles fall between 60° and 120° , to ensure skewed cells do not alter the representation. It is found that for simpler structures, [1, 2, 3] expansions often miss some of the internal topological features in the wider structure not present in the 1-cell, which are then absent in the final CPD. For this reason [2, 3, 4] expansion series are recommended at minimum where possible, which provide a reasonable trade-off between accuracy and computational feasibility.

CPDs of Known Crystal System

Relationships

The Ruddlesden-Popper Phases

The Ruddlesden-Popper phases ³¹ are composite structures with perovskite units of variable width separated by layers of rock-salt.

Let us examine how the Percifer representations vary as we consider a series of four structures which follow the general formula:

$$\text{Sr}_{n+1}\text{Ti}_n\text{O}_{3n+1}$$

For $n=0$, this gives the composition SrO, the rock-salt structure strontium oxide, with ICSD collection code 163625. This has the typical CPD for structures with periodic primitive cubic motifs. The H_0 features die once the cube forms and the space is fully connected with cycles, each of these cycles dying when the cubes faces are fully connected (Figure 23).

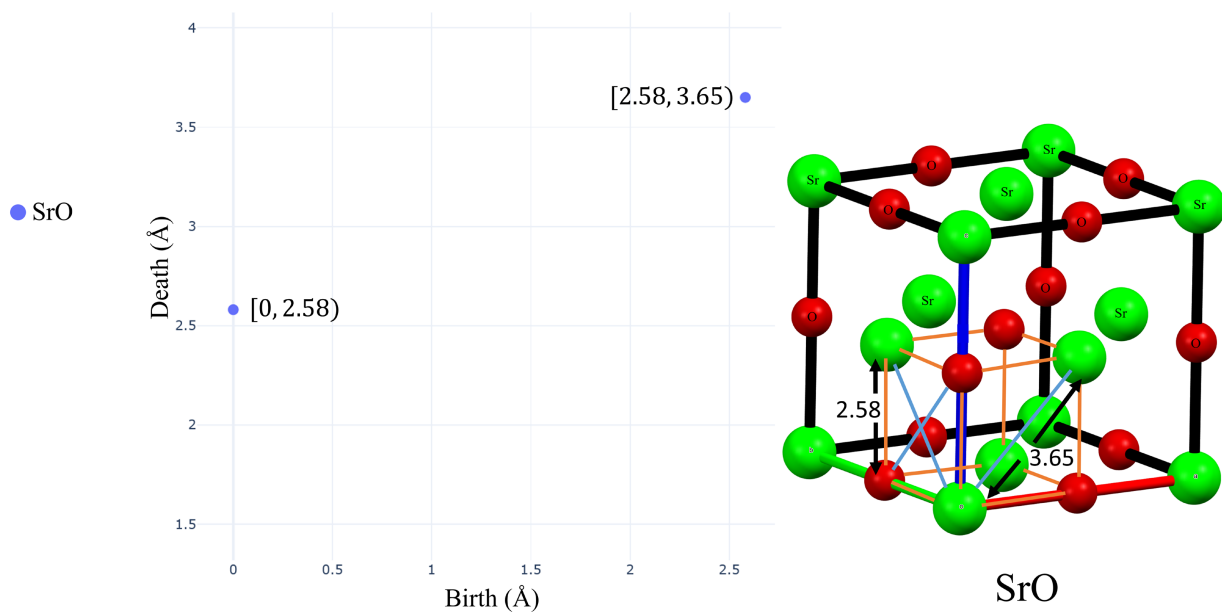


Figure 23: The SrO rock-salt CPD which has a single feature in H_0 and a single feature in H_1 shown in blue. The corresponding crystal structure is shown on the right with critical edges marked.

The next composition has a typical perovskite structure, $\text{Sr}(\text{TiO}_3)$, ICSD-34630. This structure type has a distinctive CPD consisting of two points in H_0 and a single feature in H_1 which is formed when the first feature in H_0 dies and dies with the second feature in H_0 . Unlike before, the cycle is formed between titanium and oxygen atoms, going outside the unit-cell. These cycles die when the faces of the cube are fully connected between oxygen and strontium atoms (Figure 24).

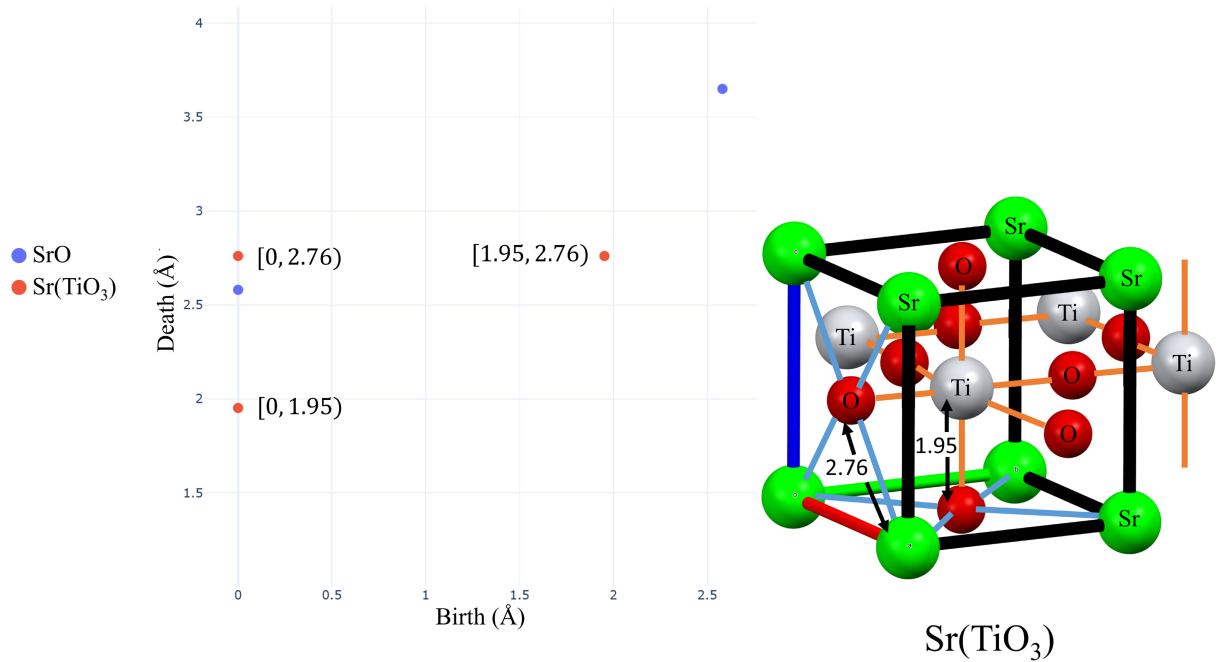


Figure 24: The $\text{Sr}(\text{TiO}_3)$ perovskite CPD shown in red which has two features in H_0 and a single feature in H_1 . The corresponding crystal structure is shown on the right, with critical edges marked.

With $n=1$ gives $\text{Sr}_2(\text{TiO}_4)$, ICSD-194713, which has alternate rock-salt and perovskite layers. Examining the CPD and comparing with the structure shows that both of these layers have corresponding points in the CPD. The slight distortion of this experimentally measured structure to the theoretically perfect structure is reflected by multiple points appearing in close proximity (Figure 25).

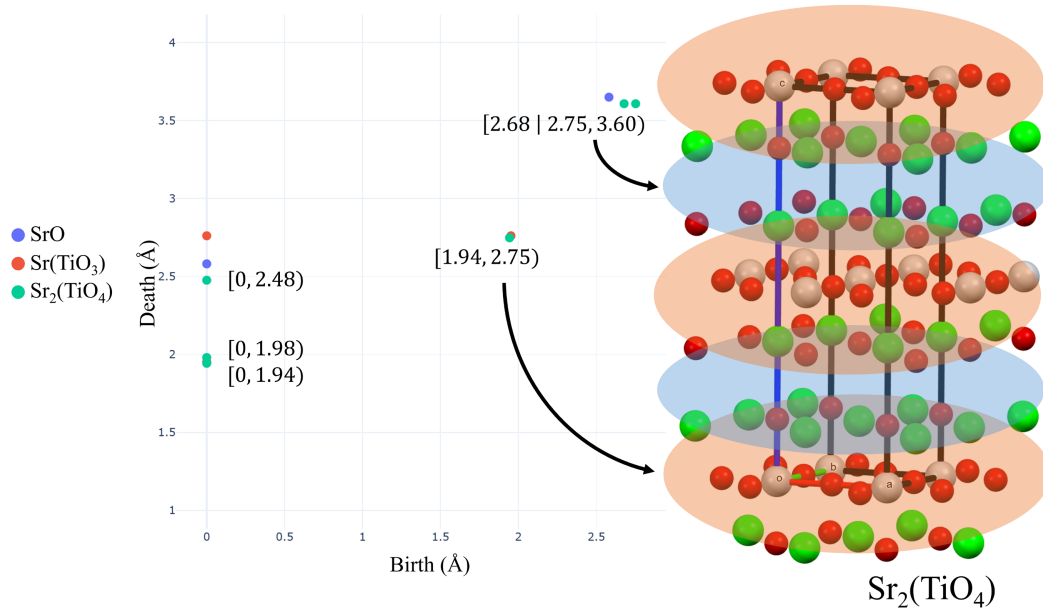


Figure 25: The $\text{Sr}_2(\text{TiO}_4)$ CPD shown in turquoise which has three features in H_0 and three features in H_1 . The corresponding crystal structure is shown on the right with the regions associated with each feature in H_1 highlighted.

We arrive at the last of the Ruddlesden-Popper phases presented here when $n=2$, $\text{Sr}_3(\text{Ti}_2\text{O}_7)$, ICSD-63704, where each of the previous layers once again have their topologies classified. Another point has appeared in a new region of the CPD, which captures the new geometry of the double perovskite layer present in this structure (Figure 26).

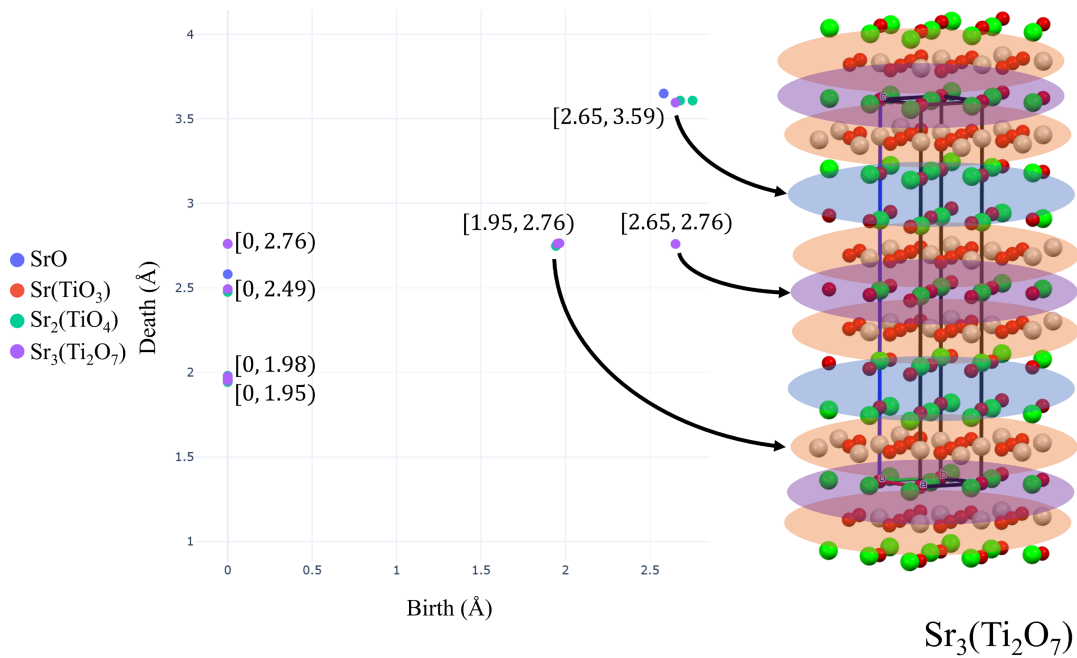


Figure 26: The $\text{Sr}_3(\text{Ti}_2\text{O}_7)$ CPD shown in purple which has four features in H_0 and three features in H_1 . The corresponding crystal structure is shown on the right with the regions associated with each feature in H_1 highlighted.

The UiO MOFs

Reticular chemistry and its applications into discovering new metal oxide frameworks (MOFs) is an active area of research. These materials consist of metal ions which are linked together with organic molecules to form large porous structures. The choice of metal and linker determines the properties of the MOF, and we would expect larger linkers to give rise to larger pore sizes. This is exemplified by the UiO MOFs³², a series of iso-reticular MOFs, with similar structures, differing only in the number of benzene rings contained in one of three linkers. There are 23 structures in the Cambridge Structural Database (CSD) of this structure type, which are named based on the linker: RUBTAK, WIZMAV, and UVAHIK.

These are large structures with hundreds to thousands of atoms in the unit cell. As such a [1, 2, 3] extension was used to construct the CPD, as it can be manually confirmed that each unit cell encompasses multiple repetitions of the periodic topology. The CPDs of each structure are processed and overlaid upon one another (Figure 27). These structures have very similar arrangements of points to one another with many characteristic features found in the same region.

A densely packed area of features on this plot are those in the green region, which are born from 1.38-1.57 and die from 2.37-2.52. Each of these points can be directly associated with a benzene ring in the linker, with longer linkers distorting the carbon sites further from the hexagonal planar structure.

The most striking correlation between this diagram and known structural relationships, is the line of points born around 2.2Å with a death time above 4Å. The birth time of this feature is dependent on the longest C-C bond found in the MOF, as this is the point when the porous cage is fully connected and its cycles may be extracted. We see that as linker size is increased, this pore diameter increases, and thus the death time of this cycle also increases.

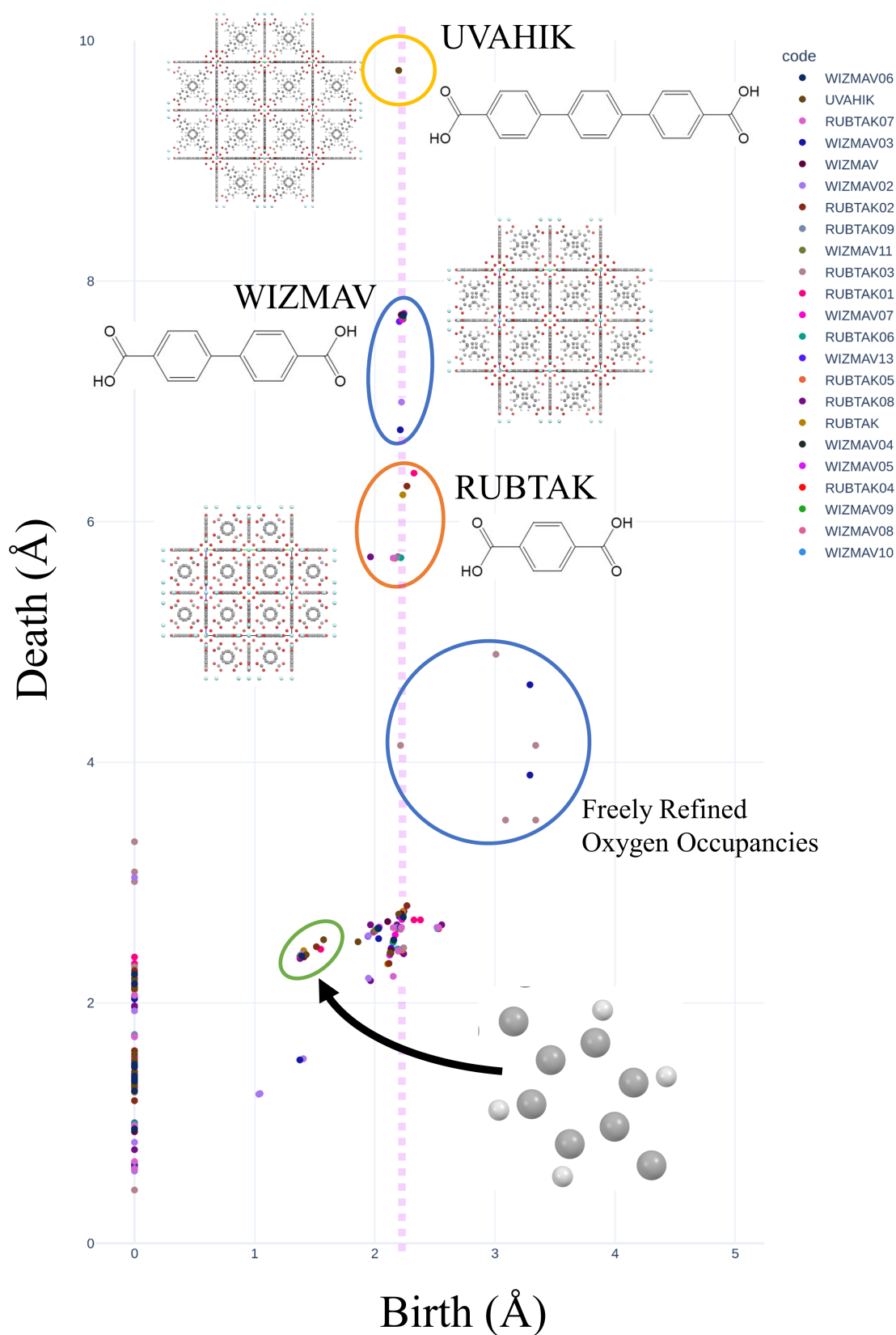


Figure 27: CPDs of the 23 materials in the CCDC with a reported UiO structure type are overlaid on one another. A dense region of points associated with the benzene rings of the organic molecules may be observed at the (birth, death) region of (1.48, 2.45). There are multiple points with a birth time of approximately 2.2 Å, which is the time each of the bonds encompassing a void connect to one another. The volume of the void increases with linker size, which is reflected by an increased death time associated with each of these structures.

Of the RUBTAK MOFs, one clear outlier can be observed, which has its porous feature die much earlier than the other RUBTAK structures. Investigating this structure, RUBTAK3, reveals that in comparison to the other RUBTAK compounds, the solvent present in RUBTAK3 has been taken into account and modelled as oxygen with freely refined occupancies, causing a smaller effective pore size. WIZMAV3 has taken a separate synthesis route to other reported WIZMAV structures, and has also taken solvent disorder into account when refining the structure. This accounts for the differences in this structure to other experimentally reported WIZMAVs, but it is also seen to have greater similarity to the RUBTAK structure with freely refined oxygen positions. Each of the outlying topological features with a birth time above 3\AA can be attributed to the additional oxygen positions in the structures of RUBTAK3 and WIZMAV3.

The Binary Compounds of the ICSD

Let us continue the investigation into binary compounds that began using ELMD (Chapter 2), with the additional insight afforded by Percifter. As visualising 3-dimensional spaces is difficult, the binary systems are close to the limit of structural complexity which may be mentally rendered. These have known trends, which allow us to validate how the Percifter metric can describe this space through reasoning.

The computational complexity of constructing a CPD means it is not possible to construct CPDs for all materials reported in the ICSD in reasonable time. Instead we focus on binary compounds, with 41,395 cif files to process into CPDs for this investigation. Due to the nature of ionic materials, materials deposited into databases be interpreted as duplicates may at first glance. In fact, of these 41,395 files, we see there are only 8,291 unique chemical compositions. Taking the set of H_0 barcodes of these structures, gives 28,907 unique barcodes. This shows that many of these compounds can be distinguished as separate crystal systems when considered up to the connected component distance. The CPD

allows us to go further than considering the distance between connected components. In particular, we can consider the set of topological features in H_1 .

Looking at the set of unique barcodes in H_1 gives an increased count of 32,216 individual barcodes. This shows that many of the structures which were indistinguishable under connected component distance are in fact separable when higher order topological features are under consideration. By combining both the sets of features contained in both H_0 and H_1 to get the combined barcode features, the number of distinctive representations again rises to 33,529. This shows that many of the structures that are indistinguishable under H_1 classification, may be separated when considering the times at which the individual components which make up each cycle connect.

Each feature from H_0 and H_1 for each CPD can be overlaid upon one another to give a sense of the distribution of topological features contained in the binary dataset under this representation (Figure 28). Like other scatter plots which contain many points, this combined representation of the space unfortunately suffers from the issue of overcrowding, making specific trends difficult to discern.

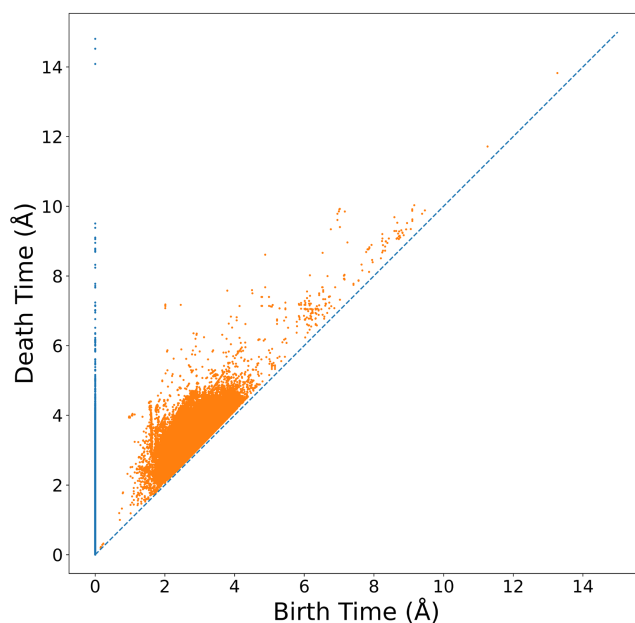


Figure 28: The CPDs of 32,216 binary crystal structures from the ICSD overlaid on one another.

A density plot of this space can be created by overlaying a pixel grid atop of this diagram, and taking count of the number of points which fall under each grid point (Figure 29), following the same methodology used to construct density maps of the ICSD with EIMD in Chapter 2. To increase visibility, the points in H_0 have been smeared to the left of the origin.

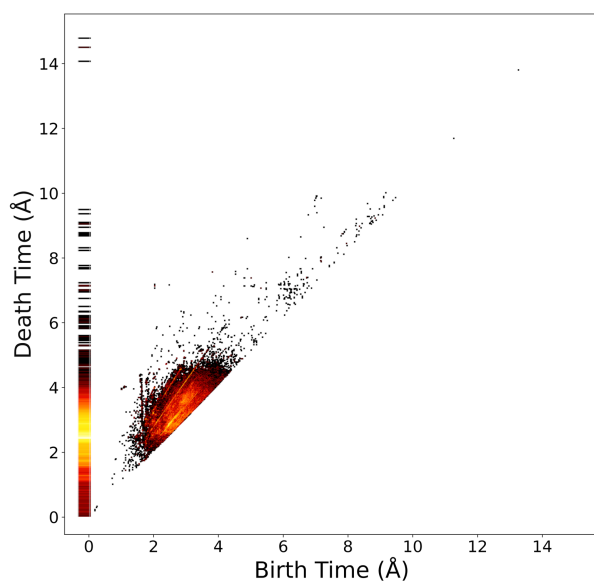


Figure 29: A density plot of Figure 28 shows the regions that are more commonly occupied by the binary system CPDs. To aid visual clarity, the single pixel line of H_0 features has been smeared to the left of the plot into the unused region with a negative birth time.

It is of little surprise that many of the connected components die in the region from 1.8\AA to 3\AA , given that we would expect the largest ionic radii found in a specific material to be in this region. Further, the distribution of lattice vectors (Figure 30) shows a peak in this region, highlighting the relative stability of this bonding distance in simple compounds.

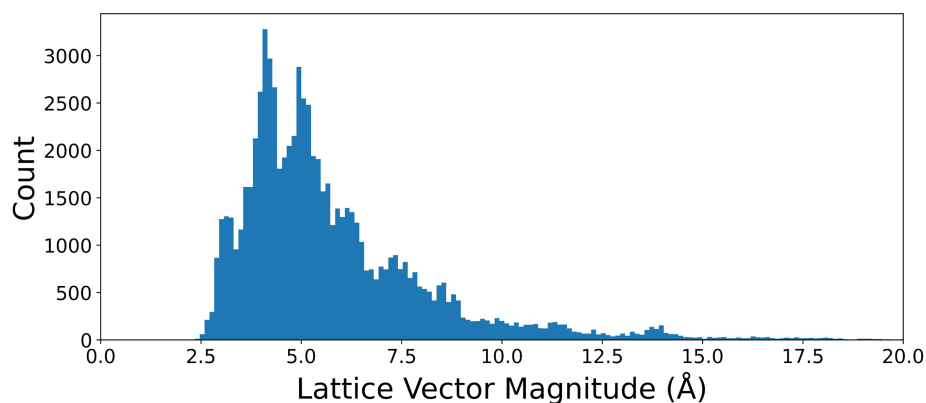


Figure 30: A histogram displaying the distribution of magnitudes in lattice vectors for the binary structures of the ICSD.

Of interest is the region of the combined persistence diagram in the H_1 region of Figure 29 where an apparent structure emerges, shown in the zoom in Figure 31.

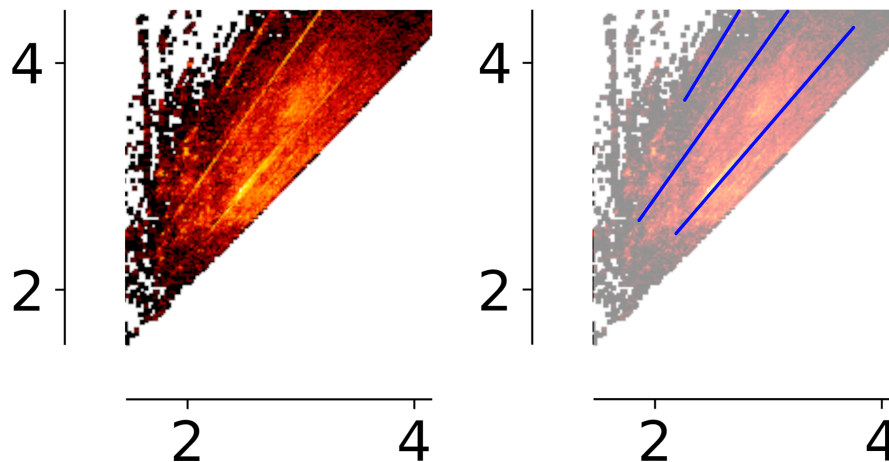


Figure 31: By expanding Figure 28 to focus on the birth/death region of 2-4Å (left), regions of greater density may be observed, which are highlighted in blue on the right.

Given the distribution of lattice vector magnitudes, we would expect a concentration of points in the birth/death radius of 2-3Å, which is the larger bright region in the centre. However, the occurrence of features falling in straight lines across the PD, highlighted in blue, was not expected. After manually extracting the associated structures through a trial and error process, it was found that every material identified along these paths is in the cubic crystal system, with a face centred symmetry group. Replotting this diagram by isolating the 8,608 FCC cubic compounds with their space groups reveals that Percifter has successfully partitioned these structures in the PD. These systems are almost exclusively found in one of the four lines of similarity, the bottom line in the diagram being two distinct segments when magnified (Figure 32).

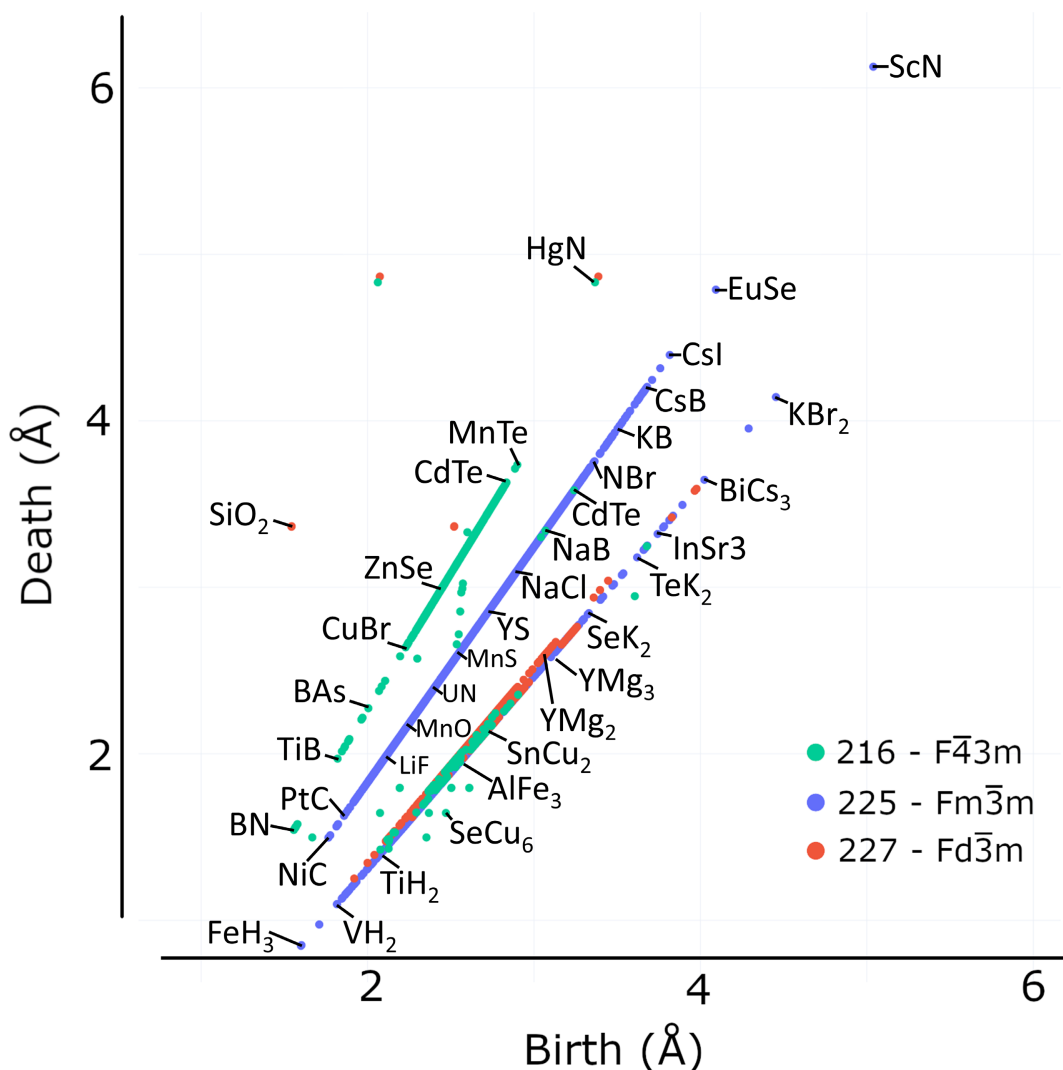


Figure 32: Extracting the 8,608 structures which fall along the regions highlighted in Figure 30 shows that each of these structures share a similarity, as they each rotate and reflect according to one of three face centred cubic symmetry operations.

Embedding Binary CPDs w.r.t the EMD

We construct a Percifier distance matrix between each of these 33,529 binary CPDs and embed this to two dimensions with UMAP, as per previous methodology. The space of 3 dimensional periodic pointclouds affords fewer natural classification methods, such as the s, p, d, f-block positions of the constituent elements in the periodic table used previously. The \log_{10} of the pointwise density can be used to highlight the changes that occur in void size across the embedding. In lieu of a continuous and consistent

labelling of periodic spaces, the Bravais groups are identified through the first two letters of the Pearsons symbol of each of the structures and used to label the plot, crudely highlighting the clustering of symmetry operations that has occurred (Figure 33).

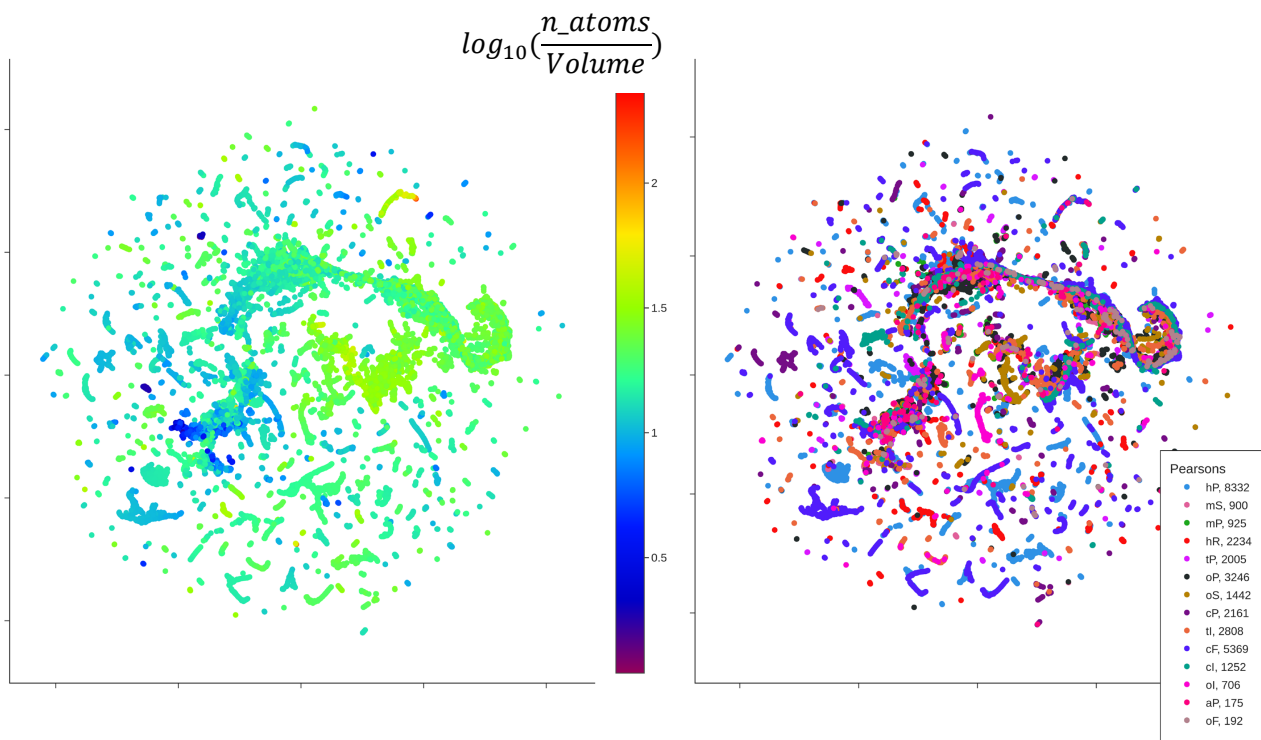


Figure 33: A Percifer distance matrix between 33,529 binary compounds of the ICSD is constructed and embedded to 2-dimensions with UMAP. The \log_{10} of the number of atoms divided by unit cell volume of each crystal (left) and pearsons symbols (right) of each material have been overlaid.

Density is shown to vary smoothly across the embedding. Distinct regions of shared space groups are seen on the periphery of the embedding. The central swirl that contains the majority of the points betrays the similarity that many of the reported binary compounds have structurally when compared with Percifer.

By concatenating the empirical formula of each cif file to its CPD, the number of unique representations grows from 33,529 unique structural representations to 37,300 combined structural and elemental representations. Despite differences in their compositional makeup, many of the binary compounds contained in this dataset have indistinguishable Percifer representations. This may be an implementation issue, as the topological features are rounded to 3 decimal places before fitting the convergent function.

If the 6,015 unique compositions are taken from this dataset and embedded to the plane via UMAP with respect to their ElMD distance (Figure 34), the resultant plot is very similar to the plot produced in Chapter 2, Figure 5. Clusters are clearly changed from the embedding presented in Chapter 2 from the greater number of compositions, however the general distribution and placement of chemistry has remained consistent. Interestingly, the variations in density remain smooth, although with greater discontinuities than in Figure 33, and the distribution of Pearsons symbols is uniformly spread across the map.

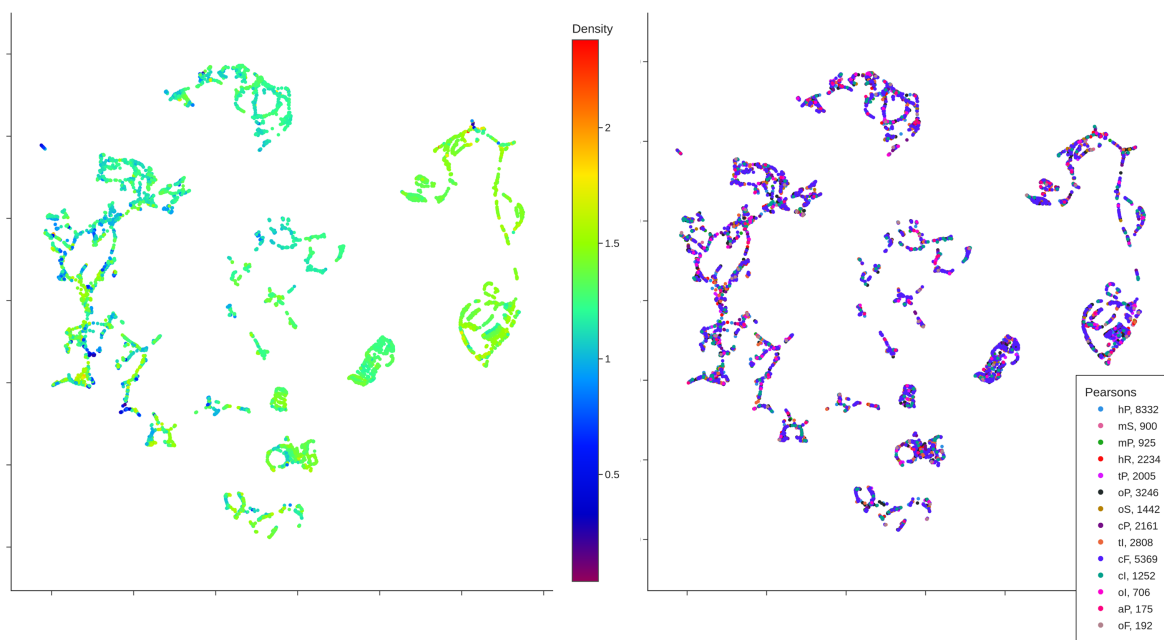


Figure 34: An ElMD distance matrix between the 6, 015 binary compounds of the ICSD is constructed and embedded to 2-dimensions with UMAP. The log10 of each of the densities (left) and pearsons symbols (right) of each material have been overlaid.

The ElMD distances may be combined with the Percifter distances. Due to the dissimilar length scales of the supporting metrics (being the difference in Pettifor number, and difference in lifetime of a topological feature), each of the distances in the matrix are scaled by the mean non-zero entries of the matrix, before the corresponding rows are summed together to give a combined distance matrix. This combined distance matrix is then embedded to the plane via UMAP (Figure 35).

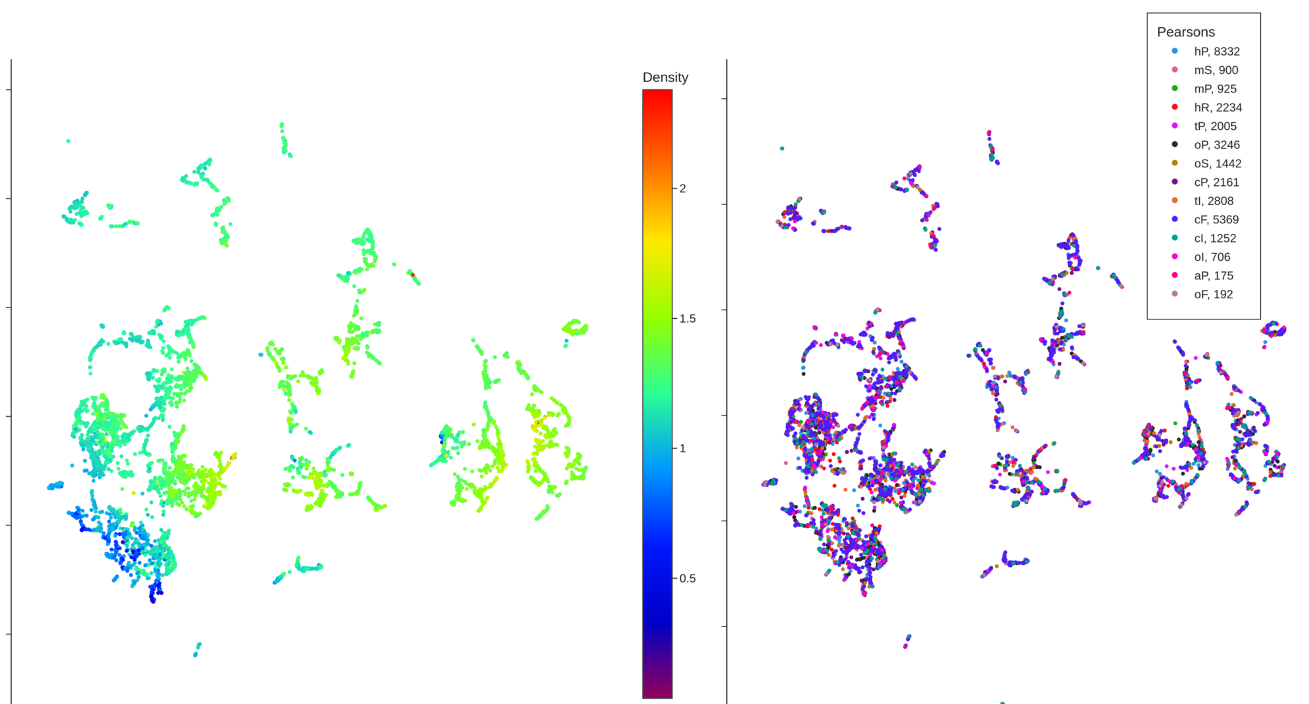


Figure 35: 6,015 compounds have their combined Percifter and EIMD distance matrix embedded to the plane via UMAP. The \log_{10} of each of the densities (left) and pearsons symbols (right) of each material have been overlaid.

It is clear from the density plot on the left of Figure 35 that there is greater structural continuity in comparison to the EIMD projection, and observing the distribution of Pearsons symbols shows a greater degree of local clustering compared to the EIMD projection, although cluster labels remain evenly distributed. The combined projection is more condensed than the EIMD projection, but remains sparser than the Percifter projection. The increased structural information has drawn together disparate regions of EIMD space that may be physically similar, but have comparatively dissimilar elements based on the modified Pettifor scale.

To judge how this has distorted the more familiar EIMD space, we shall apply the same methodology that was carried out in Chapter 2. The projection may be partitioned via DBSCAN to give chemically reasonable clusters (Figure 36, left). These same labels may then be applied to the combined EIMD and Percifter plot to see how the arrangement of points has changed (Figure 36, right).

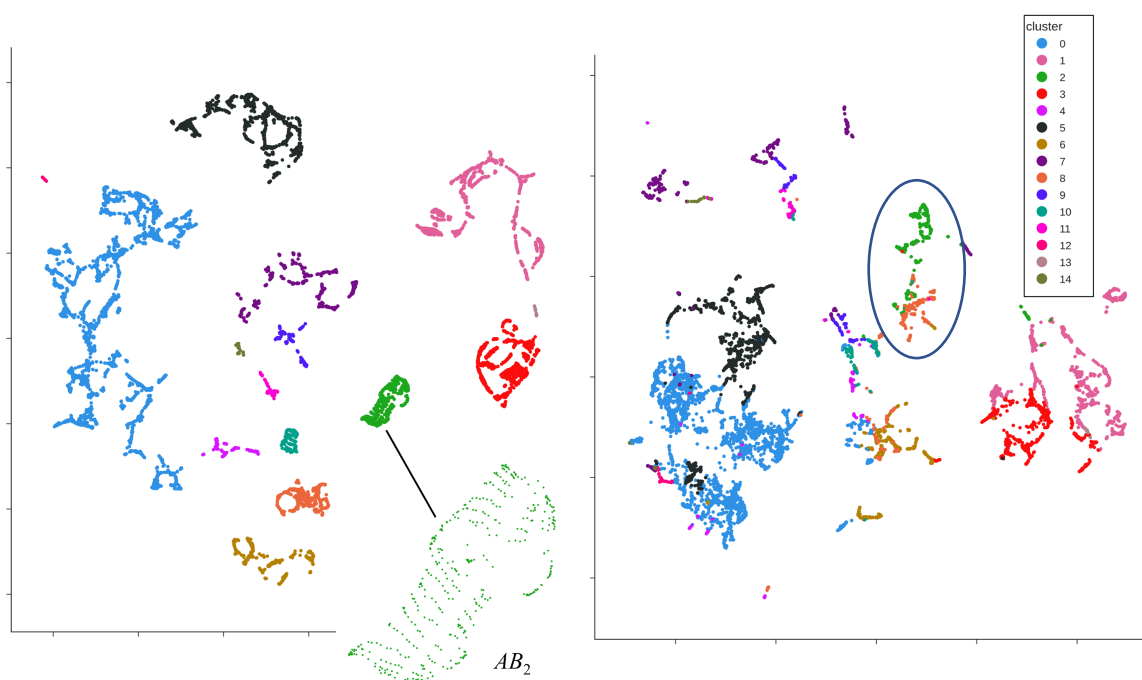


Figure 36: DBSCAN cluster labels assigned to the EIMD embedding (left) may be applied to the combined embedding (right) to highlight how this embedding has changed. A zoom of the AB_2 compounds in the blue circle of the EIMD plot is provided below.

The first observation of this combined plot is that each of the clusters have been pulled together, with less separation between distinct regions of points. Whilst the EIMD cluster labels remain in similar sections of the plot relative to one another, clusters labels merge into one another and are less clearly separated. An example where this contraction of the space has improved the consistency of observed physical trends may be found by looking at the AB_2 lanthanide-transition metal binary compounds in cluster 2. This cluster corresponds to cluster 13 of the binary compound projection of Chapter 2, and once again the AB_2 lanthanide-transition metal binary compounds have arranged themselves into separable parallel lines on the EIMD projection, shown by the zoom in the left plot of Figure 36.

In the combined projection, this cluster now forms a contiguous region with the materials from cluster 8 of the EIMD embedding, which also contain AB_2 compositions (Figure 37). The cluster 8 AB_2 compounds are characterized as having a higher electronegativity on the B ion in comparison to the cluster 2 AB_2 compounds, with the A ion still typically being a lanthanide.

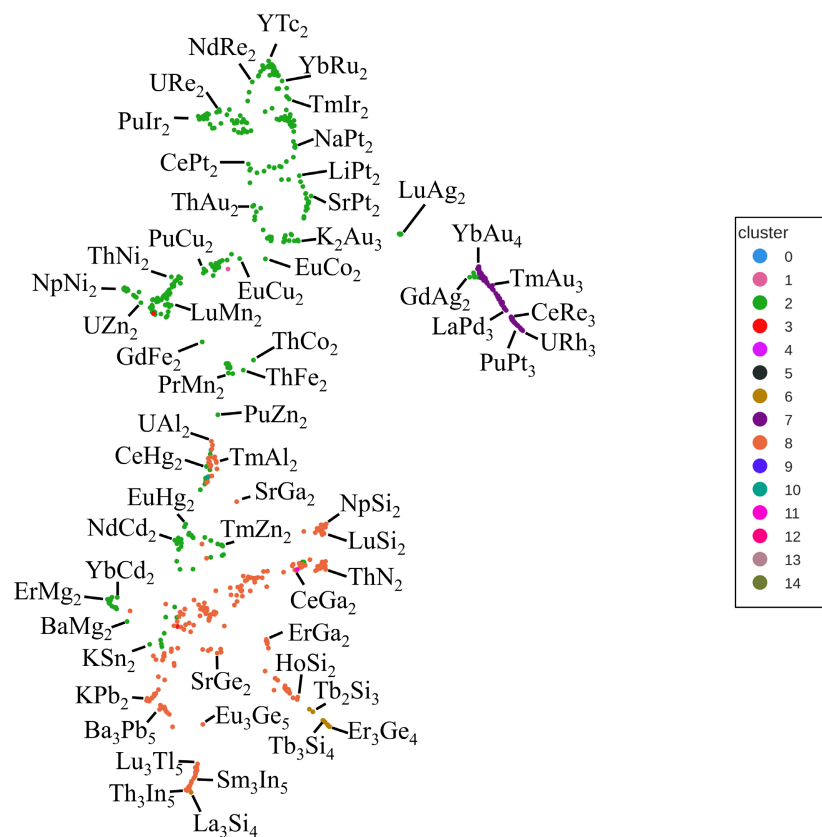


Figure 37: The AB_2 compounds investigated in Chapter 2 (green) are now associated with AB_2 compounds which are less similar compositionally, but still share similar structural characteristics.

Whilst the chemical distinction across the wider cluster can still be made out, it is clear that these structurally similar materials belong in close proximity to one another when being considered under a judgement of similarity. From this embedding we see that the addition of the Percifier distance to the EIMD distance has further tuned the alignment between the resultant metric and an experts judgement of similarity.

Topological Analysis of Solid State

Electrolytes

Each of the 403 unique room temperature compositions introduced in the experimental database of Chapter 3 was extracted and used to query the EIMTree constructed in Chapter 4. Given the affect that even minor dopants can have on resultant conductivity it was decided to only investigate those structures with an EIMD score of 0. Due to the recency in the reporting of these compounds, many are yet to be structurally characterized and deposited in an accessible materials database. Despite this, 95 unique compositions had structural files associated with them, many of the compounds returning multiple reports of the same formulation, giving 405 structural files in total. Each of these potentially different phases will be assumed to share the same conductivity for the purposes of plotting, but we shall focus on the more general trends which can be extracted.

Each of the 95 unique compositions and its associated CPD may be overlaid upon one another, as shown in Figure 38. Given the greater structural complexities found in these materials, these typically have diverse CPDs containing many points. Overlaying these diagrams allows us to visually extract trends in topological features that can be observed, which may be used to guide future discovery. With this many points, the visual density of the plots, in particular along the $x = 0$ birth time, may make extracting trends difficult. As we possess the conductivity of each material, which is the core property of interest, points have been extended along the x -axis in accordance with their $\log_{10}(\sigma)$ to draw this region further apart and further highlight trends.

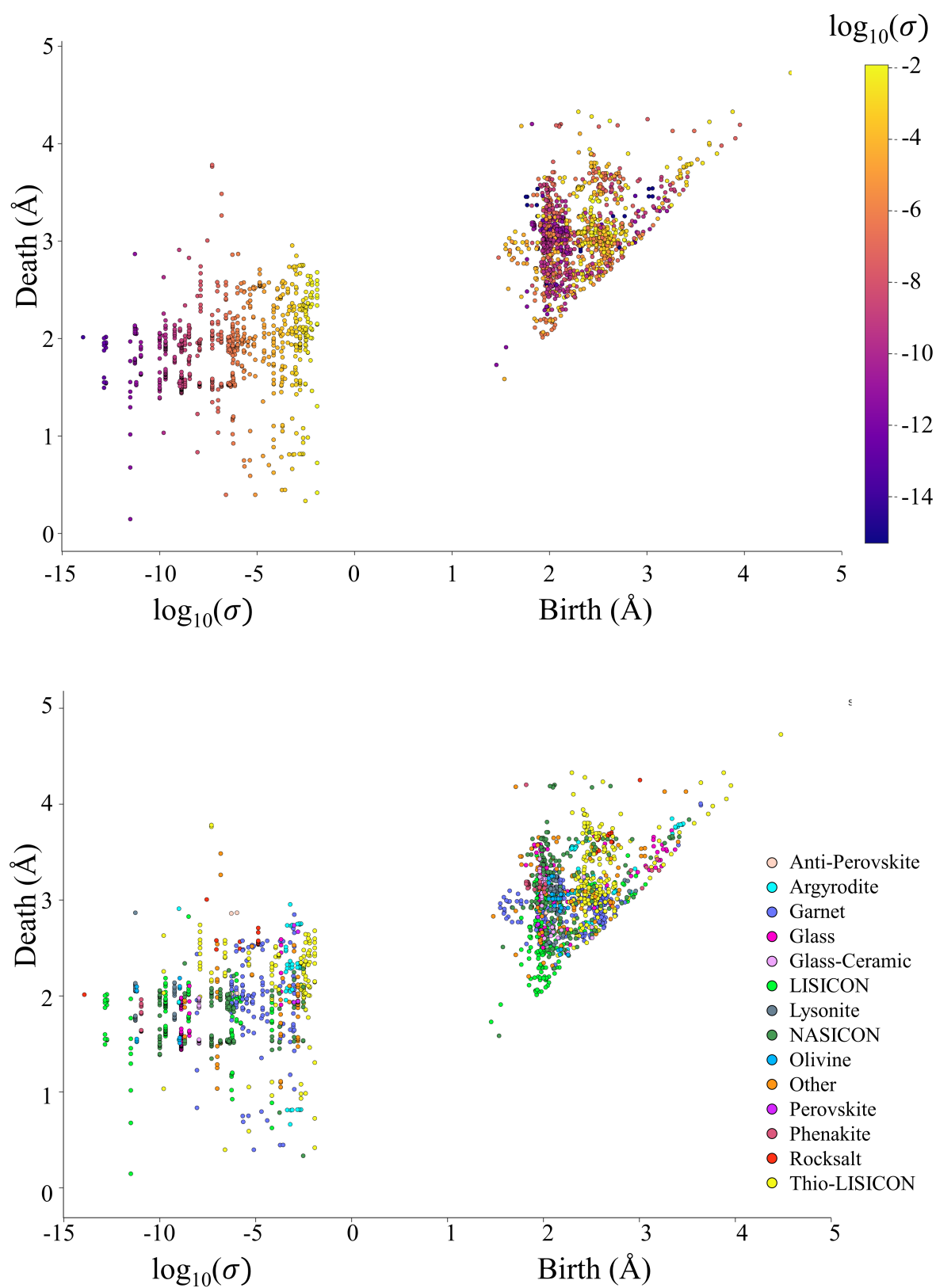


Figure 38: The CPD of 95 ionic conductors overlaid on one another. Each point is assigned a colour based on the conductivity of the material (top) or the structural family each material belongs to (bottom). To aid visual clarity, all points that fall along the $x=0$ line been smeared to the left of the plot into the unused region of negative birth time, in accordance with each materials \log_{10} ionic conductivity, σ .

Despite the limited number of structures analysed here, it is clear that natural clusters have formed between shared structural families. Although not as immediately apparent as previous case studies, examples of straight lines of doped materials can be observed as dopant concentration alters cell volume. Let us examine one of the denser regions of the plot in detail, where topological features are born between 2.51Å and 2.57Å and die between 2.95Å and 3.01Å (Figure 39). This local clustering of shared topological features includes well characterized materials such as LGPS and LLZO in addition to doped and substituted variations of these structures.

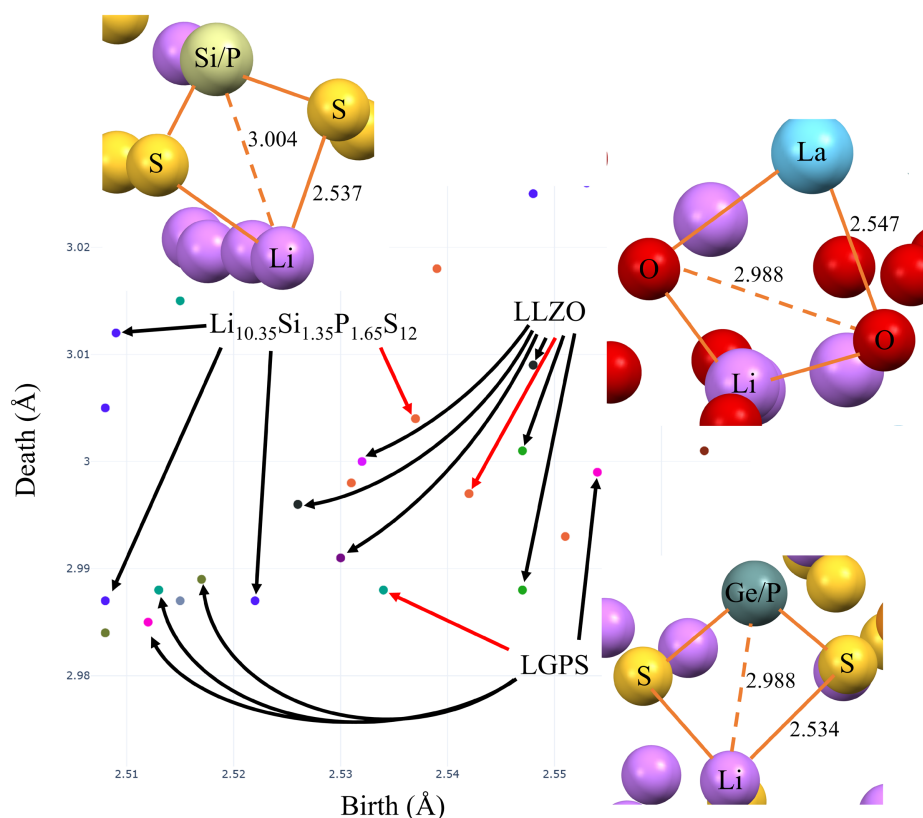


Figure 39: A zoomed portion of Figure 37, demonstrating how materials with shared physical properties occupy similar regions of the CPD.

For each of the topological features identified in this region of the combined CPD, there is a corresponding cycle that can be identified in the structure. Of note is that in each of these particular examples, the cycle contains a lithium, two atoms from the anionic

framework, and a second cation. If we look at the region around the Li there is a shared edge between Li tetrahedral/octahedral sites and another cation. Whilst this may not be a chemical feature which is a strong predictor of conductivity, it demonstrates the capability of the Percifter representation at classifying structures with shared characteristics to be in close proximity to one another, with no further embedding step required.

Another example of structural similarities arranging themselves into ordered regions may be observed by isolating the H_1 features found in the argyrodites. Here, a family of electrolytes based off $\text{Li}_6\text{PS}_5\text{Cl}$, can be found in very similar regions of the map. Doping these structures with additional silicon is seen to move the representation further away from the undoped materials. The amount of dopant appears to cause a continuous change in structural deformation, which is reflected by changes in Percifter representation (Figure 40).

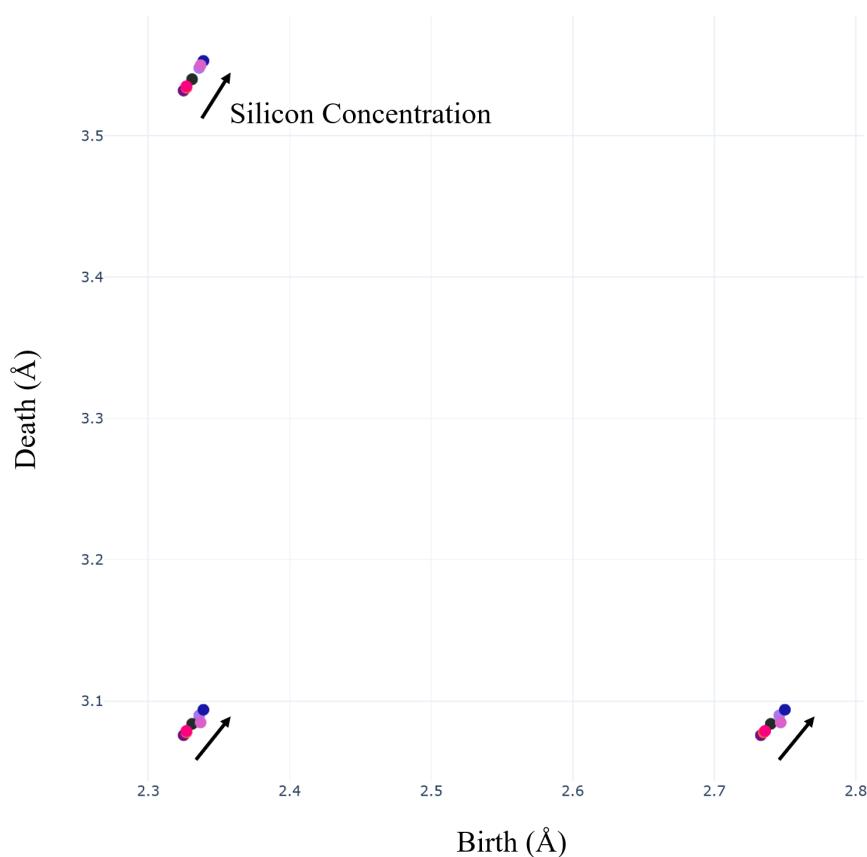


Figure 40: Variations in stoichiometry following the amount of silicon doping in $\text{Li}_6\text{PS}_5\text{Cl}$ lead to associated changes across the corresponding CPDs.

The structural topologies which make up a material lithium transport network, or those that make up the anionic backbone may be isolated and examined independently. We isolate each the identified atoms in each cif and the relational topology is investigated. When overlaying these CPDs (Figure 41), the most apparent observation is that each of the duplicate structures of each composition are found in close proximity to one another. This is unsurprising, as we would expect each of these structures to be similar, although not identical, due to tolerances within the experimental apparatus and minor differences in synthesis conditions.

Representing a material which has partial occupancy of sites with consistency is not a solved problem. In this instance, all Li sites were simply taken as a fully occupied site, and the filtration of the resultant complex taken as the representative topology of the Li positions. In actuality, fewer occupied positions would be observed in a single moment, as the Li positions are liable to change under an induced voltage. As a consequence, these representations have considerably more Li positions than they should, with many in closer proximity to one another than the rules of physics allow. This may disrupt or induce patterns and trends of the CPD, which may not actually be observed in nature. Nevertheless, one of the identified strengths of persistent homology is its robustness to noise, which is to say that pointclouds with repetitions of a point in close proximity should return similar H_1 topologies.

It can be seen that LISICONs, garnets, NASICONs, and thio-LISICONs occupy distinctive regions of the CPD, both in H_0 and H_1 persistence intervals. It is difficult to discern obvious trends in conductivity, mostly due to the highly conductive materials of the thio-LISICONs dominating the highly conductive region of the plot.

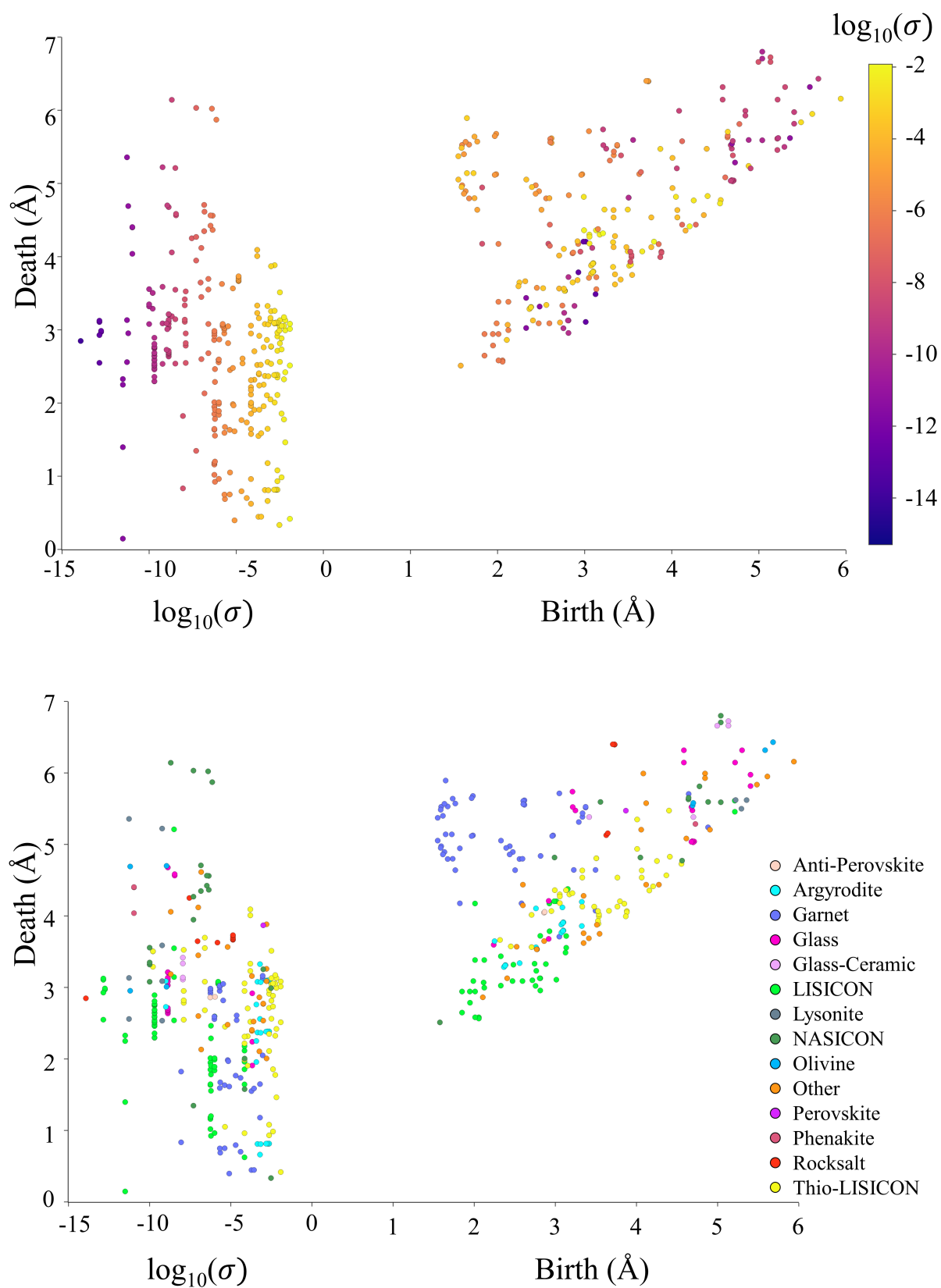


Figure 41: 95 Li ionic conductors have their anion positions isolated and the CPD calculated. We see a better separation of materials based on their structural family than when all sites are considered. To aid visual clarity, all points that fall along the $x=0$ line have been smeared to the left of the plot into the unused region of negative birth time, in accordance with each materials \log_{10} ionic conductivity, σ .

By comparison, the anionic framework of a crystal structure is often more reliably characterized. Whilst each of the anions may partially occupy sites, these are often at regular positions and, disregarding defects, at least one of the anions will always be present at these sites. This makes the co-ordinates of these elements more reliable when taking the positions of these sites from experimentally measured structures. We could assume greater stability in the representation if we only take the anion positions of the structure. What remains challenging however is assigning which of the elements in the structure are in fact anions, and which are cations.

Using an integer count of each of the elements in the composition and a list of feasible charge states for each element, a feasible assignment of charges to elements may be allocated. Elements which are negatively charged are identified and selected as the anions of the structure. This was implemented using code provided by Judith Clymo (<https://github.com/jclymo/Materials>).

Computing the anionic CPDs and overlaying these on a single plot (Figure 42) displays significantly tighter clustering than the isolated Li plot, with a substantial quantity of points found in the H_1 birth/death region of 2.8-3.2Å. Materials in the same structural families are located in similar regions to one another, however there is far greater overlap between points, making it difficult to visually discern trends. Thio-LISICONS once again have the most distinctive persistence intervals, which are located in the same region as the anion positions of the argyrodite structures.

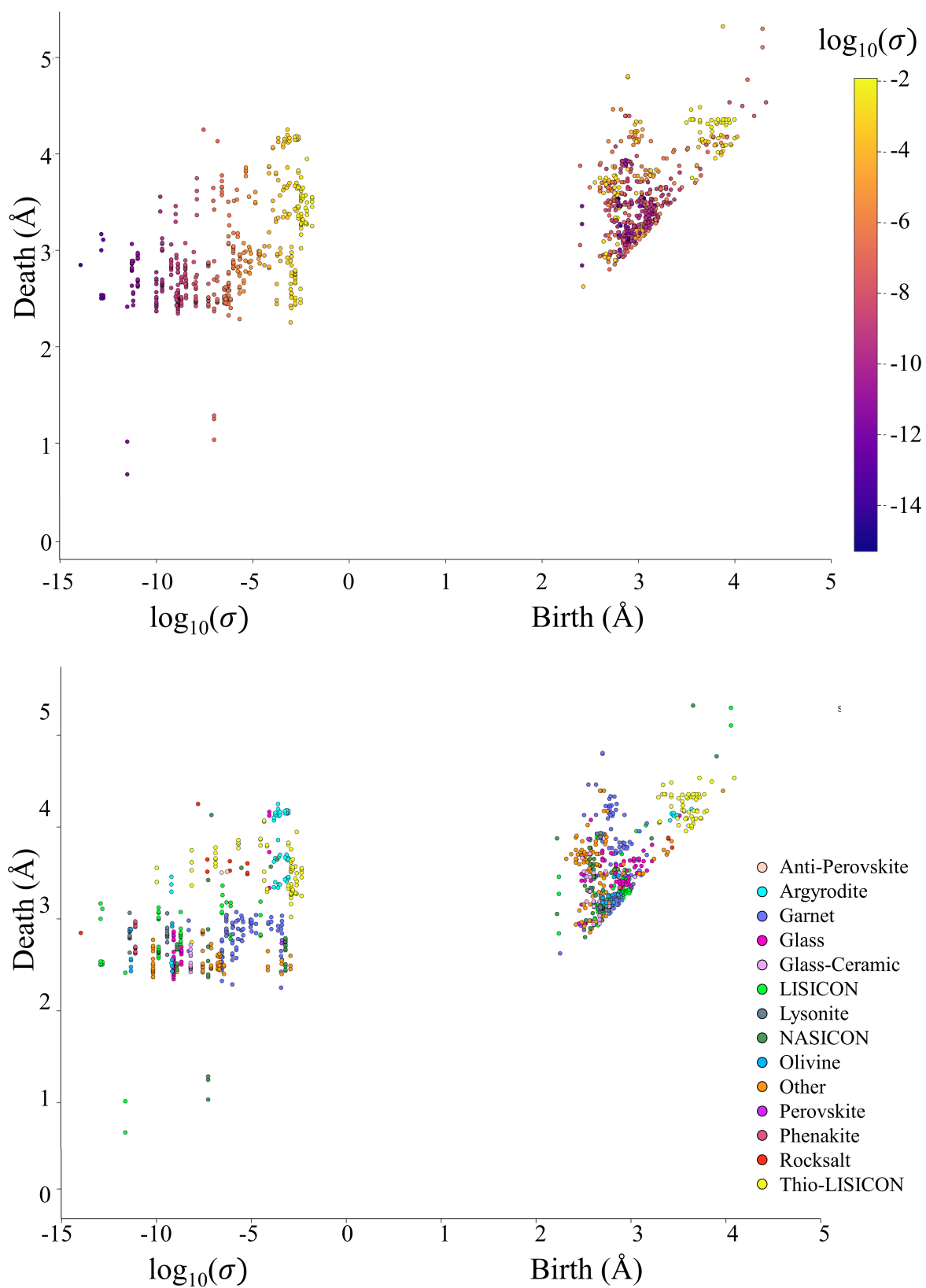


Figure 42: 95 Li ionic conductors have their Li positions isolated and their CPDs constructed. These plots contain many more points is found in the isolated anion CPD. To aid visual clarity, all points that fall along the $x=0$ line been smeared to the left of the plot into the unused region of negative birth time, in accordance with each materials \log_{10} ionic conductivity, σ .

Perciffter EMD Distances

To the practitioner, there is often a pleasing correlation between returned EMD distances and compositional similarity, which may be attributed to ingrained familiarities of the elements of the periodic table. Unfortunately the CPD, which is the underlying space from which Perciffter derives distances, does not as readily draw on learned intuition. For purposes of chemist in the loop exploration, this tool must first be calibrated against prior understanding before it can be used to support the claims of structural similarity or novelty for new materials.

Rattled NaCl

Sodium chloride has a familiar structure to crystallographers. The atoms that originally fall in a cubic structure can be “rattled” by perturbing the original lattice vectors and by altering the positions of the atoms within the unit cell. Small perturbations will lead to structures that are arguably similar to the NaCl structure, with large perturbations returning dissimilar structures.

Random samples are uniformly taken from the interval $[0, 2]$ and used as the standard deviation of a normal distributions centred around 0. The NaCl unit cell vectors have noise added to their positions according to each distribution, with atom positions also rattled by this amount. Once a structure has been rattled, it is written to a cif file, and the distance to the unrattled reference NaCl structure computed. This is repeated 20,000 times with the resultant distances plotted against the quantity of rattle (Figure 43). As a means of comparison against other metrics, the PDD EMD distance, and the ASE implementation of Oganov’s fingerprint function (OFP)³³ are additionally computed between each rattled structure and the reference structure.

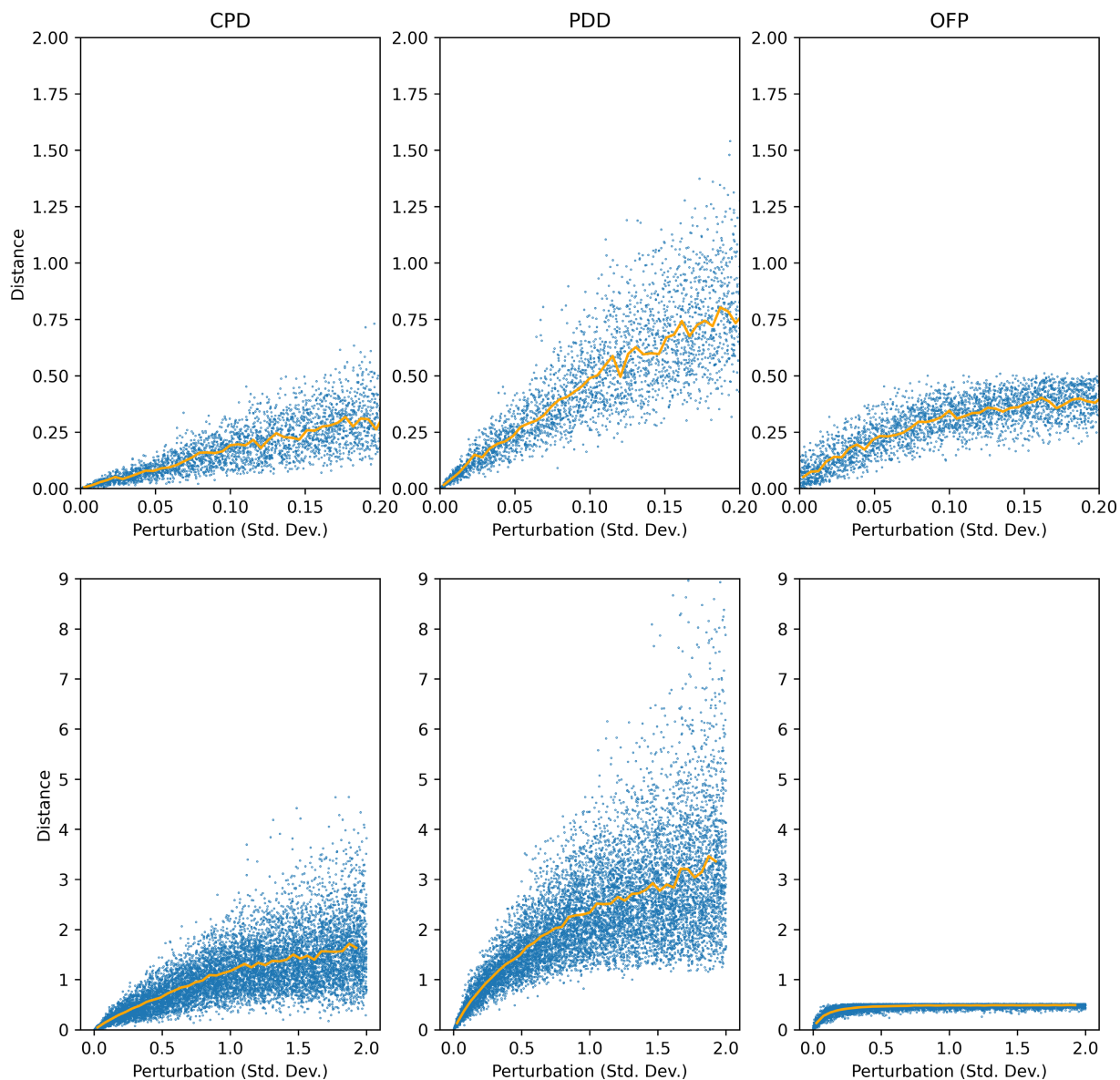


Figure 43: Variations in Percifter distance, PDD distance, and OFP distance are plotted in blue for a perturbation of 0-0.2 (top row) and 0-2.0 (bottom row). The moving average of each plot has been overlaid in orange.

Both the CPD and PDD show greater expressivity at representing the dissimilarity between more rattled structures than the OFP, which has a very flat response curve for structures with a rattle > 0.5 . It can be said, that under this metric things begin to look the same, as all dissimilar structures appear to be equally far away

from one another, despite increasingly apparent structural differences.

By comparison, the range of outputs spanned by the CPD and PDD metrics continue to grow in proportion to the perturbation of the cell. The gradient of these response curves grows sharply for smaller perturbations, before sloping off. The CPD has the more linear response curve, which is attested by the positive Pearsons R^2 score of 0.41 (Table 1). Despite possessing a negative R^2 score of -5.35, the PDD distances are clearly dependant on the quantity of perturbation. This is a shortcoming of the R^2 score, as it is unable to capture non-linear relationships reliably. Alternative measures, such as Chatterjee’s recently proposed Xi correlation ³⁴ allows this relationship to be expressed. The Xi correlation is a ranked metric which measures the strength of the monotonic relationships underlying distributions of points. Under the Xi correlation, the OFP distance is shown to have a small positive correlation with structural deviation. The PDD and CPD are both shown to have much stronger relationships, demonstrating that both of these metrics are more capable of capturing structural distortions in proportion to the quantity of distortion.

| Metric | R^2 Score | Xi Correlation |
|---------------|-------------------------------|-----------------------|
| OFP | -0.19 | 0.29 |
| PDD | -5.35 | 0.51 |
| Percifter | 0.40 | 0.50 |

Table 1: R^2 and Xi Correlation scores of the returned distance in proportion to the quantity of distortion.

Solid State Electrolytes

We take a selection of 14 electrolyte structures, which exemplify a range of materials with identified structural similarities and differences. Percifiter distances are taken between each structure, with both the H_0 and H_1 distances returned so these may be compared separately. The mean of these scores is taken to give a combined distance matrix, which is reordered to maximize the entropy along the leading diagonal to “sort” the matrix (Figure 44, top). The H_0 and H_1 distance matrices are permuted to follow the same ordering as the sorted combined matrix to allow direct comparison (Figure 44, centre and bottom respectively).

The Percifiter representation correctly partitions the oxides from the other materials. Along the diagonal, it can be seen that structurally similar materials have been clustered next to one another, despite similarity computations being executed in arbitrary order. Similar materials have correspondingly small distances from one another, with smoothly varying distances to more dissimilar structures. These distances are small in comparison to ElMD, although the focussed nature of this dataset will exaggerate that phenomenon.

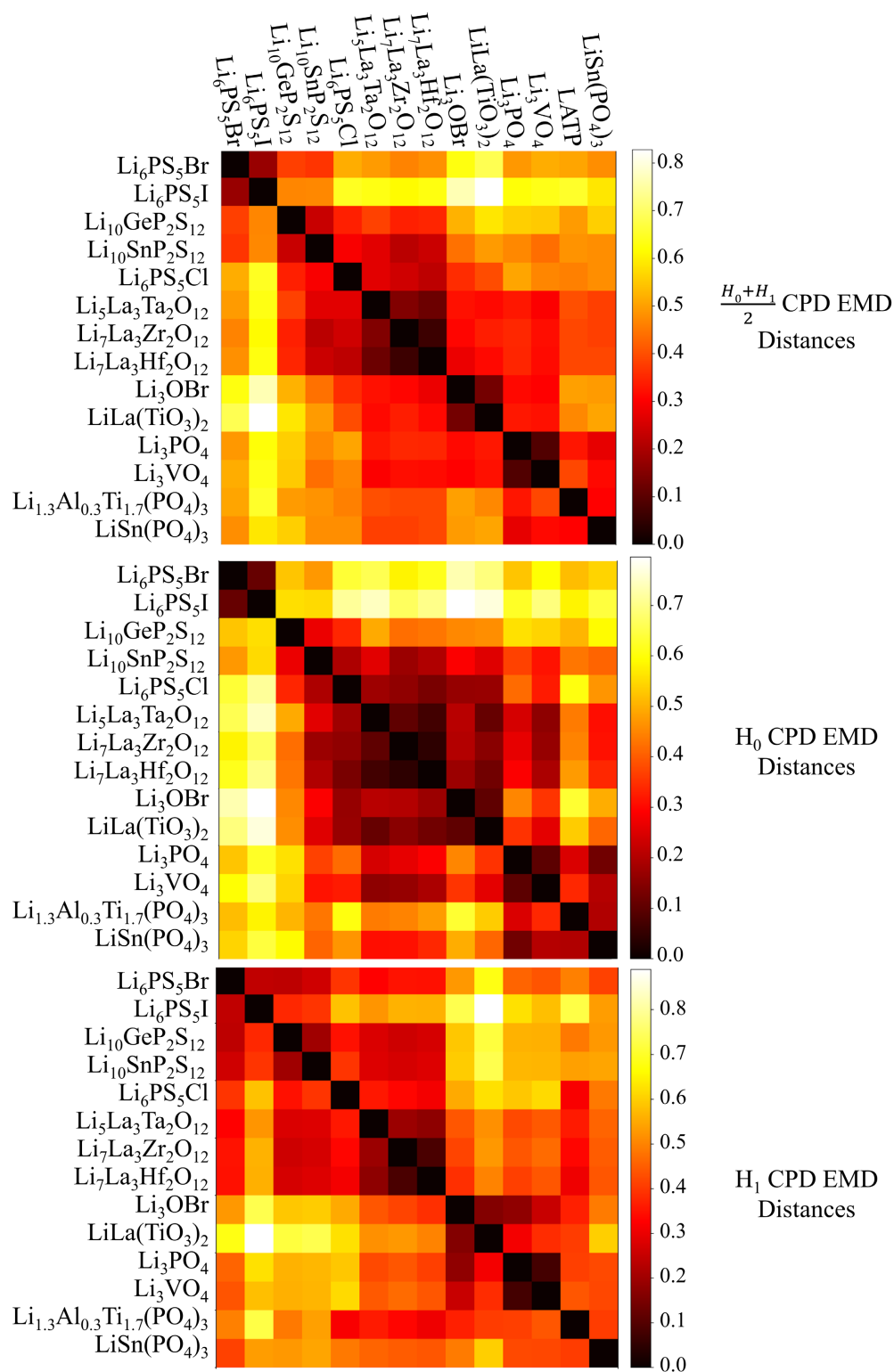


Figure 44: 14 electrolyte structures have their CPDs constructed. The combined CPD EMD distance matrix is computed and re-ordered to “sort” this dataset into regions of similarity (top). The H_0 (centre) and H_1 (bottom) distance matrices are also computed as part of this process, and are displayed using the same ordering as the combined matrix.

One outlier of interest is $\text{Li}_6\text{PS}_5\text{Cl}$, which has clustered not with the other argyrodites of the set, but instead with the thio-LISISICONs, despite known similarities to the other argyrodites, from the observed distances, $\text{Li}_6\text{PS}_5\text{Cl}$ seems to fit its position in the combined distance matrix. Examining the H_1 CPD distance matrix reveals that the distance between $\text{Li}_6\text{PS}_5\text{Cl}$ and $\text{Li}_6\text{PS}_5\text{Br}$ is lower than that of $\text{Li}_6\text{PS}_5\text{Cl}$ to its immediate neighbours, where it is observed to be an outlier of the cluster. However the great dissimilarity the $\text{Li}_6\text{PS}_5\text{Cl}$ H_0 CPD displays to the other two argyrodites in H_0 distances skews $\text{Li}_6\text{PS}_5\text{Cl}$ as being further away despite the known topological similarities.

The combined Percifter distance may be too blunt a tool to extract this kind of relationship directly, but being able to observe the underlying components allows us to appreciate how distances have been arrived upon. We may in fact go one level deeper, and plot these CPDs to directly observe the distributions of topological features and reason about whether or not these distances convey meaningful relationships. Each of the argyrodite CPDs are plotted with the LGPS CPD, both separately (Figure 45, top four plots) and together (Figure 45, bottom plot).

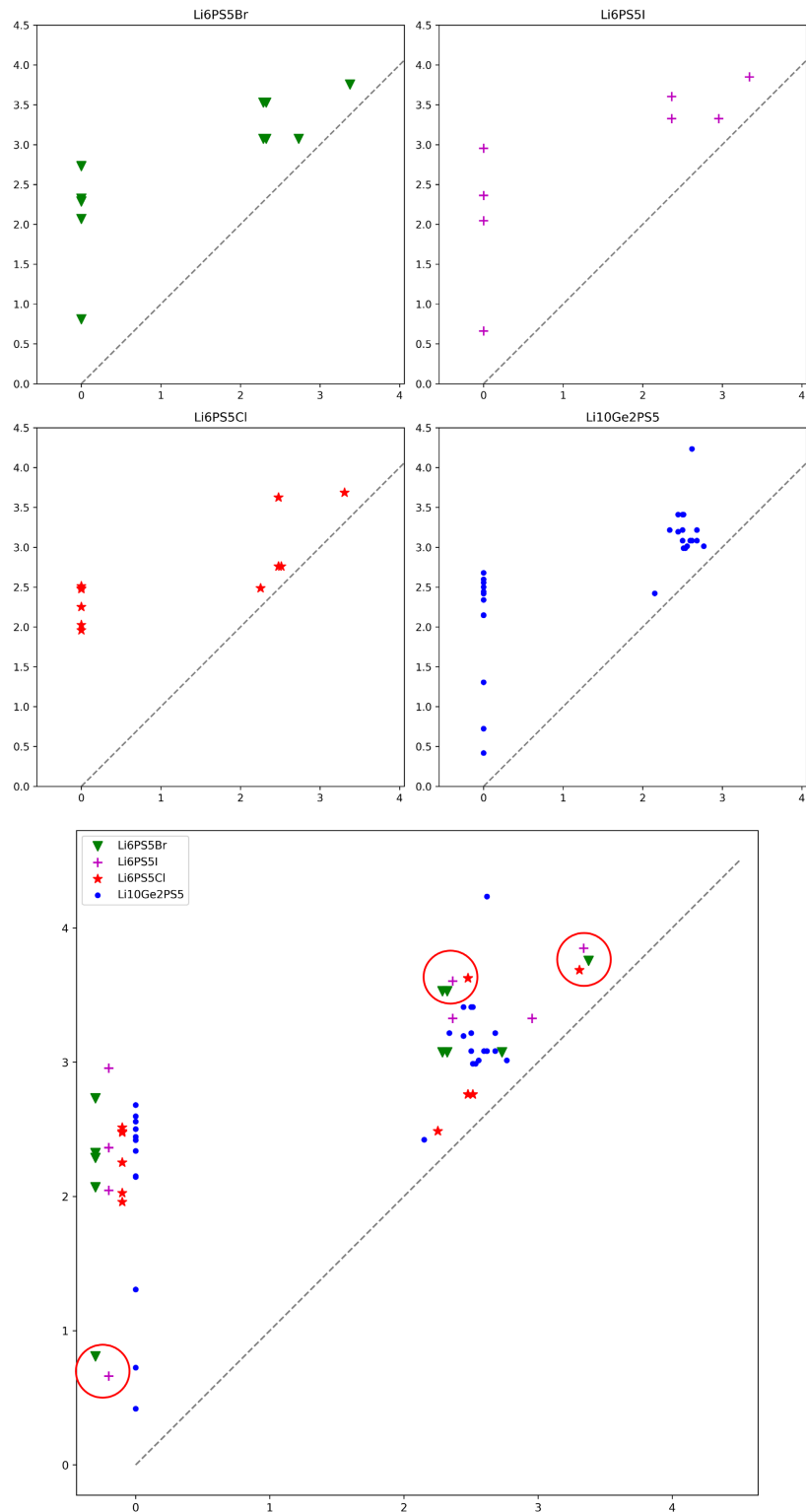


Figure 45: The CPDs of $\text{Li}_6\text{PS}_5\text{Br}$, $\text{Li}_6\text{PS}_5\text{I}$, $\text{Li}_6\text{PS}_5\text{Cl}$, and $\text{Li}_{10}\text{Ge}_2\text{P}_5\text{S}_5$, with birth and death times in Angstrom along the bottom and left axes of each plot. Each of these CPDs have been overlaid on the bottom plot to highlight the similarities between these materials (red circles) that would not be presented to the researcher using a simple metric. All points falling along the vertical $x=0$ line have been smeared to the left to prevent overcrowding.

It is immediately apparent that the $\text{Li}_6\text{PS}_5\text{Cl}$ CPD bears more resemblance to each of the other argyrodite structures than LGPS. Each CPD has 4-6 features in the H_1 set, with the H_0 features arranged into 3 clusters. By comparison, LGPS has a far more disordered crystal structure, with a majority of the features accumulated into a mass that is not particularly overlapping with the argyrodite features. These 3 argyrodites possess characterizable points in H_1 around $[2.36, 3.60)$ and $[3.34, 3.85)$, which allows these to be visually separated from one another. Quantitatively however, the outlying $\text{Li}_6\text{PS}_5\text{Cl}$ points around $[2.51, 2.76)$ carry significant mass in the CPD. As a consequence, the cost to match these points to LGPS points is lower than the cost to transport these to the argyrodite structures leading to this structural confusion.

Furthermore, in the H_0 feature set, $\text{Li}_6\text{PS}_5\text{Cl}$ does not have any features which die before 1\AA . These features are present in the other two argyrodite structures, presumably from a different modelling of the occupancy of Li positions. The actual representation given by a cif file is a fundamentally different objects to the true crystal structure, as physical constraints force experimentalists to form an approximate representation, which is an inescapable reality in many scientific disciplines.

Embeddings of the Solid State Electrolytes

The materials of the ionics database are again embedded to the plane, as demonstrated in Chapter 3, with the additional structural similarity scores given by Percifter. Each of the unique compositions of the structural dataset are first isolated, giving 95 unique materials. This gives a smaller dataset than that investigated in previous chapters, where there may be fewer example of similarity between materials to form natural clusters in an unsupervised manner. The compositions are first embedded with respect to the EIMD score between each chemical formula via UMAP, with expert assigned structural labels used to colour each point, Figure 46.

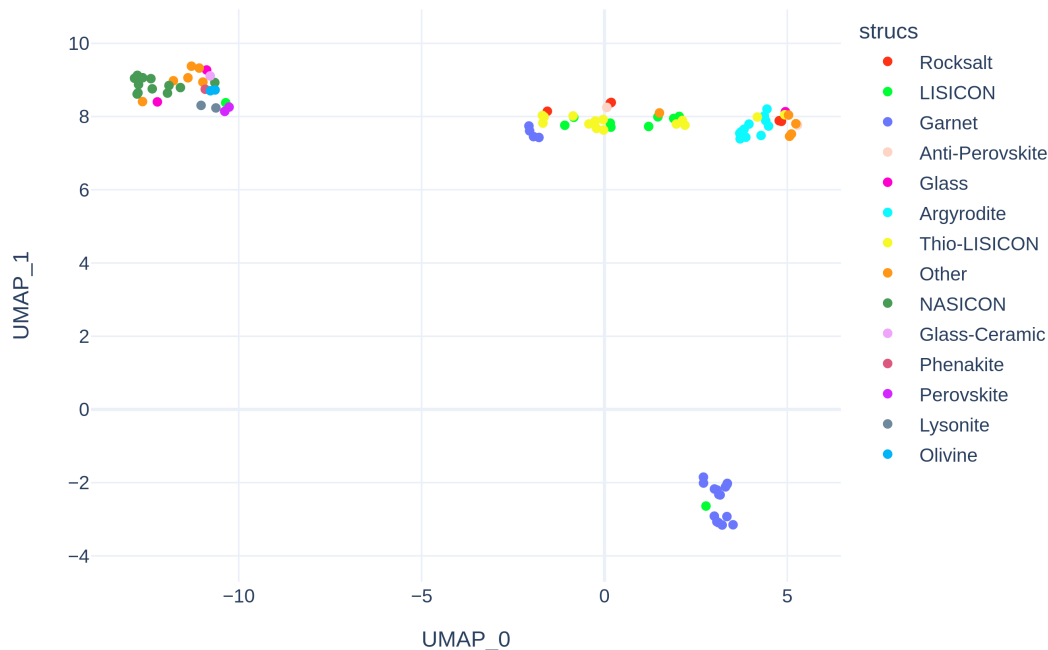


Figure 46: An EIMD embedding of the 95 materials in our structural dataset, constructed with UMAP

Each of the materials have been embedded in relatively close proximity to other materials identified as being similar. Cluster separation could be improved however. For example, the cluster found in the upper left whilst being predominantly NASICON structures, also contains a range of oxide chemistries and structure types. In the upper region of the plot, alongside some of the garnet materials (distant from the distinct cluster of garnets found in the lower right), thio-LISICONs have been interspersed with the LISICON and rock salt materials. Argyrodites are found in a single region in the upper right of the plot, but are again difficult to separate from the range of other chemistries which make up the rest of the materials clustered in this region.

Next, all chemical information is disregarded and just structural positions are examined. The CPD of each material is taken using all atomic sites without any elemental positions isolated. A distance matrix of Percifter similarity scores is constructed

between the CPDs of 95 unique cif files by summing together each of the H_0 and H_1 EMD distances. This distance matrix is then embedded to the plane using UMAP, with structural labels overlaid, Figure 47.

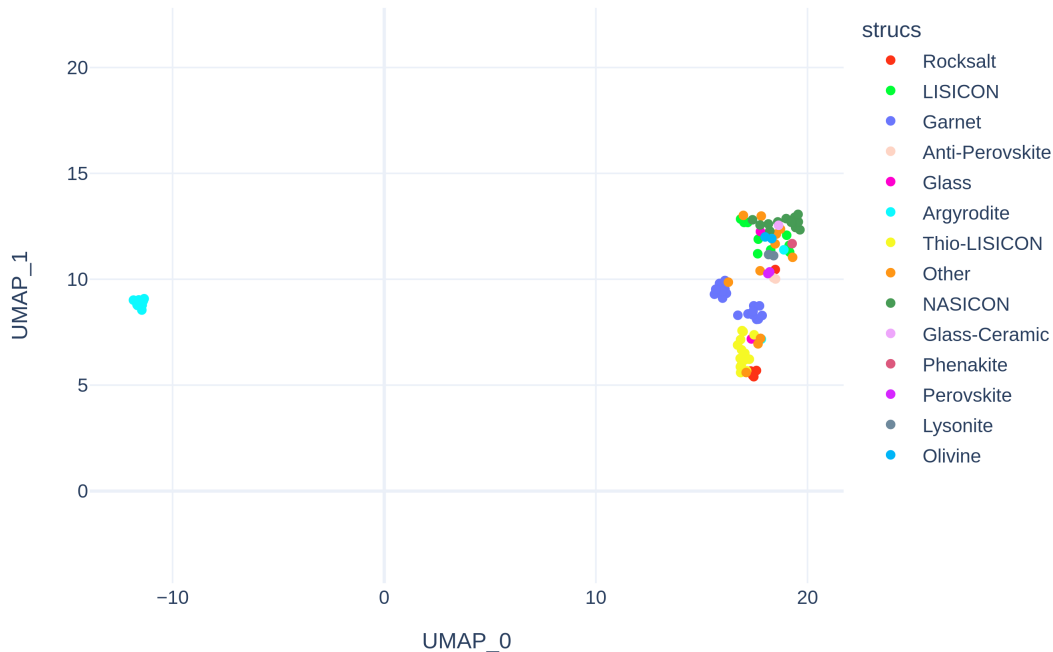


Figure 47: A Percifter distance embedding of the 95 materials in our structural dataset, constructed with UMAP

The immediate observation is how well separated each of the Br and I anion argyrodite structures have been partitioned from the other materials in the cluster on the left. The remaining structures are found in the rightmost cluster, in a much denser arrangement of points than is found using compositional information. Despite the dense clustering, the materials have been well ordered by structural family, with distinctive areas of shared structural characteristics. Boundaries between these areas may be too blurred to extract using automated techniques, but an alignment of structural similarity (and argyrodite dissimilarity) is demonstrated by the metric.

As noted previously, the supporting metrics for each similarity measure are on different orders of magnitude and thus EIMD scores would be over represented by simply summing these

together. The ElMD and Percifter distance matrices are scaled to have a mean value of 1 by dividing each value in the matrix by the respective mean value of that matrix. These two matrices are then combined by summing together their values, with the resultant matrix embedded via UMAP, Figure 48.

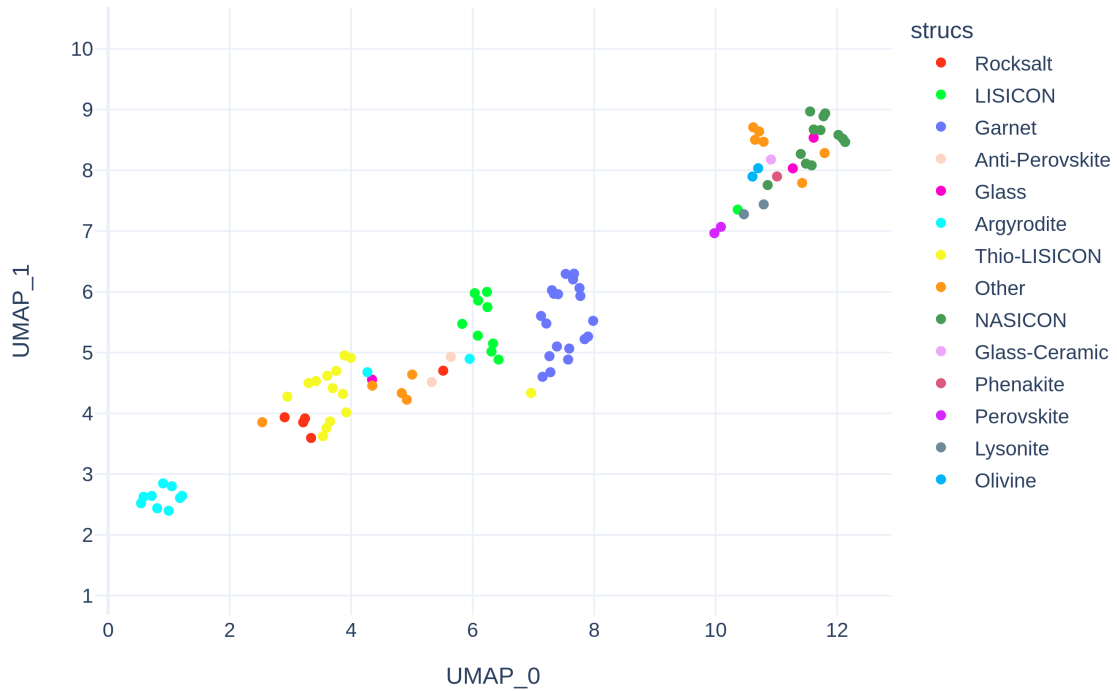


Figure 48: A combined ElMD and Percifter distance embedding of the 95 materials in our structural dataset, constructed with UMAP

In comparison to the previous two embeddings, the clusters of Figure 47 are well separated from one another. Argyrodite dissimilarity has still been captured, but individual materials may now be distinguished from one another. Each of the remaining structural labels are found in close proximity to one another, with the “Other” and “Glass” labels being the only identified clusters spread across the embedding. Thio-LISICONs, LISICONs and garnets are in similar, yet distinct, regions of the embedding, each separable from their neighbours.

The similarity of the NASICONs to the remaining oxide materials, previously highlighted by each metric when considered in isolation, remains present in this embedding. By contrast however, each of these materials have now been drawn further from one another, allowing them to be differentiated, and making chemical trends simpler to follow.

Further Work

One of the underlying principles that statistical analysis relies upon is the assumption that there is some underlying structure to the data that is being modelled. There has been much research into how to represent data to exploit this assumption, and many of the step change advances in neural network performance have come not from improvements to the densely connected neural network itself, but in the data representation stages. Convolution operations, attention networks and message passing graph operations all attempt to permute the initial data to make this underlying structure between training points more apparent to the final neural network stage.

From this observation we believe the CPD may be a candidate for further investigation in data driven materials discovery. There are existing methods to represent a persistence diagram for a neural network (PersLay³⁵), and it is possible this may be a reasonable starting point to manipulate the CPD for a richer representation of a material to be input into a neural network.

One of the strengths of the representation is its ability to display trends in a tangible format by comparing CPDs. Automatically extracting these similarities would allow more of these trends to be explored. At present, correlations have been extracted manually by cross referencing against the original crystal structure. Whilst EMD between CPDs generally follows structural patterns, we can see by the counterexamples provided that this does not always pair a structure to one we may think of as similar. This may be a

drawback of the metric, and other methods of deriving similarity may have stronger alignment with chemists judgement.

It should be noted that a similar approach has been investigated by Teresa Heiss and Herbert Edelsbrunner, which has not been published but was presented in the Applied Algebraic Topology Research Network (AATRN) Online Seminar, September 22, 2021³⁶. The core improvement this approach takes over the work presented here is the choice of function that will track the multiplicity of each feature as unit cells are expanded to infinity, that we will find the limit of. In this chapter, while fitting the function $f(x) = \pm \frac{a}{x-b} + c$ has shown the principle can work on each of the examples, it is likely that this could lead to imprecise final ratios, especially for more complicated motifs. In comparison, Heiss looks to fit a separate polynomial function for each persistence feature. The use of these polynomials will surely trace the limit of each feature as the unit cell grows in size to greater accuracy than the $f(x)$ used in this work.

Conclusions

Here we have presented an overview of the Percifter topological representation of periodic structures. This is preceded by the mathematical prerequisites at an introductory level, followed by an overview of the core Percifter theory based on the convergence of sequences of expanded unit cells. The Percifter representation is fully plottable, giving an expressive planar representation of the periodic structure called the convergent persistence diagram (CPD). Similarities in periodic features may be reasoned and derived from overlaying and comparing CPDs and cross referencing against the original crystal structures.

A more quantitative method of comparison between CPDs is the earth movers distance (EMD), where this metric may be applied in a range of unsupervised techniques as in previous chapters. We demonstrate how this may be used in combination with compositional similarity measures to give combined chemo-

structural similarity scores. This combined score is demonstrated to express known chemical trends in the small data regime.

Widening the arsenal of tools available for cheminformaticians to interrogate their datasets allows us to unearth more of the previously unnoticed quirks. We hope that these techniques inspire others to further explore the trends present in materials datasets and to advance the tools we use to represent these spaces.

References

1. Choudhary, K. *et al.* Recent advances and applications of deep learning methods in materials science. *Npj Comput. Mater.* **8**, 1–26 (2022).
2. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput. Mater.* **5**, 83 (2019).
3. Langer, M. F., Goeßmann, A. & Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *Npj Comput. Mater.* **8**, 1–14 (2022).
4. Carleo, G. *et al.* Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).
5. Roth, R. S. Classification of perovskite and other ABO₃-type compounds. *J. Res. Natl. Bur. Stand.* **58**, 75 (1957).
6. Sendek, A. D. *et al.* Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
7. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys Rev B Condens Matter Mater Phys* **87**, 184115 (2013).
8. Himanen, L. *et al.* Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
9. Deringer, V. L. *et al.* Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
10. Helfrecht, B. A., Semino, R., Pireddu, G., Auerbach, S. M. & Ceriotti, M. A new kind of atlas of zeolite building blocks. *J. Chem. Phys.* **151**, 154112 (2019).
11. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M. & Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Netw.* **20**, 61–80 (2009).
12. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
13. Blatov, V. A. & Proserpio, D. M. Periodic-Graph Approaches in Crystal Structure Prediction. in *Modern Methods of Crystal*

- Structure Prediction* 1–28 (John Wiley & Sons, Ltd, 2010). doi:10.1002/9783527632831.ch1.
14. Choudhary, K. & DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *Npj Comput. Mater.* **7**, 1–8 (2021).
 15. Cheng, J., Zhang, C. & Dong, L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater.* **2**, 1–11 (2021).
 16. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How Powerful are Graph Neural Networks? *undefined* (2019).
 17. Widdowson, D. & Kurlin, V. Pointwise distance distributions of periodic point sets. Preprint at <https://doi.org/10.48550/arXiv.2108.04798> (2022).
 18. Rabadán, R. & Blumberg, A. J. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. (2019).
 19. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 454–463 (2000). doi:10.1109/SFCS.2000.892133.
 20. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P. & Harrington, H. A. A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**, 17 (2017).
 21. Hatcher, A. *Algebraic Geometry*. (2005).
 22. Munkres, J. R. *Elements Of Algebraic Topology*. (CRC Press, 2019). doi:10.1201/9780429493911.
 23. Vietoris, L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Math. Ann.* **97**, 454–472 (1927).
 24. Edelsbrunner, H. & Harer, J. *Computational Topology: An Introduction*. (American Mathematical Soc., 2010).
 25. Jiang, Y. *et al.* Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *Npj Comput. Mater.* **7**, 1–8 (2021).
 26. Mohamad Moosavi, S., Xu, H., Chen, L., I. Cooper, A. & Smit, B. Geometric landscapes for material discovery within energy–structure–function maps. *Chem. Sci.* **11**, 5423–5433 (2020).
 27. Niggli, P. Krystallographische und strukturtheoretische Grundbegriffe (Handbuch der Experimentalphysik, Bd VII, 1). *Geol. Fören. Stockh. Förh.* **51**, 121–122 (1929).
 28. Delone, B. Neue Darstellung der geometrischen Kristallographie. *Z. Für Krist.-Cryst. Mater.* **84**, 109–149 (1933).
 29. Bauer, U. Ripser: efficient computation of Vietoris–Rips persistence barcodes. *J. Appl. Comput. Topol.* **5**, 391–423 (2021).
 30. Zhang, S., Xiao, M. & Wang, H. GPU-Accelerated Computation of Vietoris–Rips Persistence Barcodes. Preprint at <http://arxiv.org/abs/2003.07989> (2020).
 31. Ruddlesden, S. N. & Popper, P. The compound Sr₃Ti₂O₇ and its structure. *Acta Crystallogr.* **11**, 54–55 (1958).

32. Cavka, J. H. *et al.* A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. *J. Am. Chem. Soc.* **130**, 13850–13851 (2008).
33. Valle, M. & Oganov, A. R. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallogr. A* **66**, 507–517 (2010).
34. Chatterjee, S. A new coefficient of correlation. Preprint at <http://arxiv.org/abs/1909.10140> (2020).
35. Carriere, M. *et al.* PersLay: A Neural Network Layer for Persistence Diagrams and New Graph Topological Signatures. in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 2786–2796 (PMLR, 2020).
36. Heiss, T. Geometry and Topology of Periodic Point Sets, for example Crystals. *AATRN Online Seminar*. Available at <https://youtu.be/LiGivu0maD8>, accessed 19/05/2023, (2021).

Conclusions and Further Work

The discovery of new solid state Li-ion electrolytes for next generation energy storage devices remains an ongoing pursuit. Despite reports of materials which advance many of the desired properties, a suitably hard and stable material with low electrical conductivity and high ionic diffusivity remains elusive. This multi-decade search for these materials leaves a rich legacy of incremental advances from which we may correlate trends to direct future approaches, and further our understanding of these systems. Recognising and characterising patterns in data is a fundamental approach to scientific discovery. Evolutionary pressures have tuned this aspect of cognition, and uncovering a previously unnoticed trend often leads to involuntary feelings of elation, accompanied by renewed passion to continue an unyielding analysis.

Unfortunately, the huge number of previously reported experiments coupled with the inherent complexity of representing a material means that uncovering such trends by reviewing the literature is no simple task. We believe the tools presented in this thesis are of benefit to the materials researcher, by condensing and highlighting such patterns. In science, we do not get our explanations for free, these must be deduced. But, knowing that a trend exists in the first place, is the first step in identifying why such a trend exists.

The Element Movers Distance

A core part of this thesis has investigated novel methods to quantify similarities between solid state compounds. The element movers distance (EIMD) is first introduced as a composition based measure of similarity. By observing known elemental similarities, the judgement of compositional similarity may be reframed into an optimal transport problem. This is demonstrated to give a more sensitive measure of comparison than the commonly used

compositional Euclidean distance (CED), as the CED may only consider the similarity of elements which are shared between two compositions.

The ElMD is shown to be a consistent numerical measure between two compositions which has allows us to embed and partition datasets of materials with automated methods, letting us verify the alignment of these distances with known chemical relationships. Whilst this is a computationally intensive procedure when compared to simpler metrics, modern implementations allow this to be calculated in reasonable time for use in high-throughput applications.

This has been primarily applied to materials datasets with unsupervised machine learning. One of the most direct methods of observing the efficacy of these metrics is by constructing embeddings of materials datasets with respect to the metric. Humans are adept pattern recognisers in two dimensions, and projecting a high dimensional space onto the plane often allows us to identify trends and inconsistencies in the dataset by eye. This does not tell us anything about why such trends exist. However, uncovering a trend in a dataset allows us to apply our knowledge and begin the process of reasoning as to why such a trend may exist.

Carrying out a DBSCAN clustering on a UMAP embedding of a materials dataset w.r.t. the ElMD has shown success at partitioning compositional spaces into chemically reasonable subgroups. These cluster labels allow us to manually verify the efficacy of the metric, as chemically similar materials are found in the same clusters. This also provides the usual benefits of unsupervised labelling, which allows us to make basic assumptions about new materials based on the cluster in which they are located.

We demonstrate how these cluster labels may be used to judge model performance in supervised learning as an alternative to k -folds cross-validation. A form of data leakage can be said to occur when data in the testing set carries enough similarity to samples in

the training data that the model does not attempt to form a new prediction, but simply returns the known value of the data found in the training set. By partitioning a dataset into holdout sets for leave one cluster out (LOCO) cross-validation, we may train a model on the majority of our data before testing its predictive performance on a chemically distinct holdout set. This allows us to judge whether a model has truly learnt to extrapolate the relationships found in the training set to new chemistries. As a core aim of data driven materials discovery is the exploration of entirely novel chemical domains, we believe judging models based on their LOCO performance gives a more truthful measure of performance than the more favourable results that will be obtained through k -folds cross-validation.

The integration of these metrics with supervised or generative ML models has not been explored in this thesis. Nevertheless, the EIMD has been applied by others in a range of ML workflows¹⁻⁷, often used to judge a hypothetical materials similarity to previously reported materials as a proxy measure of either feasibility of synthesis or novelty. This is a core component of the Descending from Stochastic Clustering Variance Regression (DiSCoVeR)⁵ approach by Baird *et. al.*. This technique improves on the embeddings presented in this thesis, by taking each of the materials in identified clusters and replotting these with the property of interest against the local density of each composition in the EIMD space. In doing this, a trade-off between a materials predicted functionality and compositional novelty may be established, which allows the practitioner to make reasoned judgements on which candidate formulations should be prioritised for synthesis.

In the past decade, advances in deep generative models have improved to the point that they may no longer be dismissed as mere curiosities. In particular, the wide adoption of diffusion models^{8,9} harks the beginning of a new paradigm in visual design which may well prove to be as revolutionary as the invention of the camera. It is known that optimal transport can be used with

generative adversarial networks (GANs) to improve stability during training, leading to higher quality outputs ¹⁰. More recently optimal transport has been integrated into generative methods for materials ¹¹, and it is our belief that further integrating the chemically aligned metrics presented in this thesis could lead to improvements in quality for other generative workflows.

Throughout these investigations it has been assumed that the modified Pettifor number ¹² is a consistent measure of elemental similarity. However chemical inconsistencies must be introduced when compressing a space as information rich as the known elements into a linear scale. For example, the assignment of magnesium at index 73 brings it in close proximity to iron, manganese, and zinc, placing it far away from calcium, at index 16, which does not align with known chemical understanding. Whilst additional chemical scales have been included as part of the ElMD package, a thorough investigation into the optimal elemental embeddings to extract domain knowledge has not been carried forward.

Monotonic integer linear scales bring clear advantages. These are fundamentally more intuitive, as the number line carries a strong mental association with physical distance that is not easily afforded to us with higher dimensional and real valued scales. Further, these scales allow us to use much faster algorithms for computing EMD, which brings many practical benefits for high throughput applications. Nevertheless, it may be the case that using more complex elemental descriptors will give a stronger alignment between the abstract measure of chemical similarity as judged by the chemist, and the numeric measure of similarity returned by the metric.

ML-Driven Design of Ionic

Conductors

The initial aim of the study presented here was the presentation of the Liverpool ionics dataset, accompanied by an investigation of the performance of state of the art ML models at predicting conductivity from composition. Due to the limited size of previously available datasets, a new dataset was constructed. The new dataset was an order of magnitude greater than previously available, with 403 unique compositions with their recorded A.C. impedance spectroscopy at room temperature, and 790 readings in total. These readings were taken from the literature, spanning a conductivity range of $5.00 \times 10^{-16} \text{ S cm}^{-1}$ to $2.50 \times 10^{-2} \text{ S cm}^{-1}$. Each of the compositions was assigned a structure type during initial data gathering, with 15 structural families covered in total. This took the efforts of 30 researchers to validate each of the entries due to the wide discrepancies in how these properties are reported in the literature requiring multiple validation stages.

The timescale of this investigation was lengthened by the covid-19 epidemic. Learning effective methods of remote communication and adjusting to new means of working was not an overnight process. Whilst decreasing the individual burden, having a large team also increases the administrative burden of synchronising schedules and circulating detailed instructions. Project milestones must take other team members prior responsibilities into consideration, assigning more time to ensure that the request of validating data is a reasonable request for all parties. To increase the user experience, a bespoke interface was developed presenting each researcher with a single entry and the relevant information to validate each recorded conductivity.

Once validated, this dataset was interrogated with recent unsupervised and supervised ML tools. Structural representations were not used, as only composition was recorded when gathering data. Embeddings of the compositional space allow us to observe

the chemical diversity in the given dataset. Using EIM2D we construct compositional maps of the ICSD with this dataset, and of the dataset in isolation. In the combined embedding we see that the majority of the compositions are found in localised regions of the map, suggesting there may be areas of unsurveyed potential that could be targeted for future exploration. In the embedding containing only compositions of the dataset, similar structure types fall in close proximity to one another in isolated clusters. This enables the use of the clustering algorithm DBSCAN¹³ to automatically assign labels to compositions which are similar to one another under the EIMD.

In big data regimes it can often be assumed that the training set is a representative sample of the global population, and that new testing data will also be drawn from the same distribution. Thus, k -folds cross validation may be applied as a reasonable measure of how models will perform in applied scenarios. In materials discovery we know that the reported materials form a limited subset of all possible formulations, which has been anthropogenically biased through human prioritisation. Furthermore, materials of interest are often interesting due to being dissimilar to previously reported materials, *i.e.* out of distribution.

The 9 cluster labels identified with DBSCAN may be used to isolate compositions of chemical similarity from one another, and thus be used in leave one cluster out cross validation (LOCO-CV), by training 9 supervised models on a majority of the dataset and testing the predictions on each left out cluster in turn. LOCO-CV gives a more realistic judgement of how our models perform when exploring new phase fields, where the model has not been exposed to these interacting chemistries before.

Regression and classification models that can accurately predict a materials conductivity from its composition are of clear benefit. These would enable us to screen wider regions of compositional space for exciting formulations before the lengthy process of synthesis. State of the art random forest and deep learning

techniques were used in conjunction with the two cross validation schemes discussed above, with the models performance at forming predictions on the chosen test sets taken as the final measure of performance.

In each of the cross-validation schemes and experimental hold out set, we see that the ML models typically outperform the control studies. The best performing deep learning classification model predicts whether a material will have a conductivity greater or less than $10^{-4} \text{ } \sigma/\text{cm}^{-1}$ with an accuracy of 0.71 and an MCC of 0.38 on LOCO holdout sets. This model has improved its score over the control studies performance, such that we believe it can be used in further work to screen novel phase fields as a preliminary step prior to more computationally demanding screening procedures. These models are imperfect, and do not replace the judgement of a trained chemist. They may however be applied by trained chemists to increase the mental bandwidth they can dedicate to effectively search for materials of interest, by first discarding a greater number of uninteresting materials. False negatives will mean some formulations that deserve investigation will be overlooked, but one hopes that the volume of materials that are worthy of further investigation will be increased overall.

False negatives are to be expected when using a lossy descriptor such as the chemical composition. Ionic conductivity is a factor of multiple physical phenomena interacting in a many body 3-dimensional space, and thus it is unsurprising that this model fails to capture all the irregularities of this complex system when given so little information about each material. We believe for preliminary screening phase fields before deeper investigation this is an acceptable trade off. Given the large quantities of data that are reported each year, future investigations should incorporate structure as an additional screening step once a crystal structure has been determined computationally, as it is demonstrated that ML models with these features consistently outperform composition only models.

Percifter

Persistent homology (PH) was proposed whilst investigating potential methods of representing the structures of ion-conducting materials. PH marks the birth and death times of topological features, which is to say the number of connected components and cycles, of a space as points in a 0-dimensional pointcloud are expanded into spheres of growing volume. This has been explored in the literature by other groups to represent materials ¹⁴, but the periodicity of crystalline structures ensures that the PH representation of a finite pointcloud contains noise from the fragmented structural features imposed by periodic boundary conditions. The utility of PH is not hindered despite these inaccuracies, and it is demonstrated to be a useful tool in materials discovery ¹⁵. Practical methods to reduce this noise have been reported in the literature, with large numbers of atoms in the unit cell ¹⁶ and atom specific PH ¹⁷ proposed as methods to better represent the topology from the initial cif file.

In this chapter we introduce a novel approach to reduce the impact of these boundary features, Percifter, based on a sequence of increasing unit cell expansions. The ratios of topological features have a converging function fitted, and their limit at an infinite expansion taken as the final ratio in the convergent persistence diagram (CPD). Boundary features should have a negligible or negative final ratio, and thus may be culled from the representation. A similar approach has more recently been proposed by Teresa Heiss at the applied algebraic topology seminar “Geometry and Topology of Periodic Point Sets, for example Crystals” (<https://youtu.be/LiGivu0maD8>), which should attain greater accuracy in final CPD representation than currently returned by Percifter, but no preprint is available online at present.

In our work it is demonstrated that various structural trends shared by materials can be identified by following patterns across multiple convergent persistence diagrams (CPDs) by eye. The birth and death time of the connected components and cycles present in the

periodic structure discards information about much of the longer range and higher order interactions, which could be correlated with properties of interest. However, by focussing on specific features, what is presented to the researcher is a more digestible subset of space which allows the most salient trends to be extracted manually. This has been demonstrated by several examples.

For the Sr-Ti-O Ruddlesden-Popper phases, the separate rock salt, perovskite, and double stacked perovskite layers may each be associated with distinct features on each CPD, which are found in similar regions across the separate CPDs. We show that trends in void sizes may be shown across series of MOFs, by the UiO UVAHIK, WIZMAV, and RUBTAK structures. These materials have shared similarity, as these are isostructural with linker size increasing by a single benzene ring as we go through the series. This shared similarity is reflected in the CPD, as the death time of points on the CPD increase in proportion to the void size.

Binary structures of the ICSD were revisited after the EIMD chapter. As there are many thousands of reported structures composed of two elements in the ICSD, extracting structural trends through direct observation is a more challenging task.

Nevertheless, each of the structures with high symmetry cubic lattices possess characteristic pointsets on the CPD, allowing these structures with shared features to be isolated from the remainder of the dataset.

We show that by treating the fractionally weighted intervals of the CPD as piles of earth with mass in proportion to the ratio of the interval, the earth movers distance may again be applied, this time between two CPDs. This is demonstrated to align with chemical judgement, and is shown to separate collections of structures by their perceived structural similarity, and embeddings of such sets with respect to this metric appears to cluster points with similar motifs in close proximity, respecting the symmetry group of the crystals. The differences returned by this metric span a smaller range of values when compared to the EIMD, due to the smaller

range of values that are covered by the supporting space, as most points in a CPD are typically found in dense regions between 1-5Å whereas elements on an integer scale will range from 1 to 118. When the Percifter metric is used in isolation this difference in scale has no effect on other tasks, but when combined with other metrics this requires rescaling to ensure that the structural contributions provided by this metric are not overwhelmed by the magnitude of others.

The Percifter metric can be combined with the ElMD, but due to the aforementioned differences in underlying length scales, we apply a scaling operation first. A complete distance matrix of both distance scores is first computed, before each matrix is divided by the respective means. The rescaled matrix will carry the same variations of the metric, but with a new mean of 1. Thus, both distance matrices may be summed together to give a combined chemo-structural distance matrix. This may be used as a lookup table for the combined distance between two materials in further downstream tasks. In our combined embeddings it is shown that inconsistencies which are present when judged solely on elemental similarity can be resolved when additional structural information is included.

It is possible that, like humans, supervised ML models find the relationships between chemo-structural space and properties of interest simpler to follow using CPD descriptors than other descriptors. These are not however in the fixed length vector format necessary for most downstream ML tasks, and whilst multiple methods of embedding a PD into a fixed length vector exist^{18,19}, a canonical representation is yet to emerge. The CPD adds further complexity, as fixed length representations in the literature will assume integer, not fractional, counts of each persistence interval in the PD. Thus, new fixed length representations must be defined to encode CPDs into a suitable format for supervised ML tasks.

We have not explored any methods of reversing the process and generating crystal structures from their representative CPD in this thesis. The inverse design of a pointcloud from its persistence diagram (PD) is partially solved in the finite case²⁰, by iteratively updating the positions of a pointcloud until the generated pointclouds PD is in agreement with the target PD. However, PDs are not invariant descriptors of pointclouds²¹, and as different pointclouds can generate the same PD, these remain imperfect tools.

For the examples given in the Percifter chapter, each of the atomic sites were treated as solid periodic structures. In many areas of crystallography this is a reasonable assumption, but various compounds exhibit irregularities through the existence of mobile ions and point defects. This means it may not be possible to define a canonical representation of such a crystal structure, as separate samples will have slightly different distributions of ion positions, leading to differing fractional occupancies for each of the refined positions. Further, minor fluctuations of an external electric field will cause the positions of mobile ions to migrate leading to altered, although similar, periodic structures.

This simple observation has been disregarded in this work. The cif parsing library that is used in the Percifter implementation (atomic simulation environment) does not record fractional occupancy, and as such all structures are assumed to have full occupancy for all sites in the cif. This means that the processed CPD is not truly representative of the typical crystal structure, as there will be significantly more mobile ions included in the finite representation than are truly found in a finite region of the periodic crystal.

A simple workaround would be to generate a collection of representations for a given structure, with the probability of selecting a mobile position to include in each representation decided upon based on its occupancy. The resultant CPDs could be pooled together, and for a suitably large number of representations of the initial structure, this may give a more faithful representation

of the true structure. However, this still has the potential to generate deceiving representations.

In a real crystal, overlapping sites with partial occupancy could not physically contain ions at both sites simultaneously, but encoding this restriction is not trivial and this situation could occur with the above approach. Whilst methods of characterising dynamic/temporal graphs have been formalised²², these assume known, or observable, evolutions of the edgeset and pointset. We are not aware of such techniques for pointsets where points have stochastically weighted visibility.

Closing Remarks

If we could not judge distance we could not avoid tigers. The trait of repeatedly making the correct judgement of this intrinsic property of nature runs far back in our genetic tree. This simple feature of human cognition is well appreciated by mathematicians and good geometric reasoning is often valued as a trait of strong logical deduction. Applying this skill to unimaginably complex systems, such as the space of feasible materials, requires tangible representations that give the psyche something to grasp on to. When we make predictions for thousands of candidate materials, these must be presented to the practitioner in some manner. An unending sheet of data is a gauntlet which cannot be attacked. A map is a concise representation which fosters curiosity.

The ease that modern deep learning frameworks take to set up means they no longer require heavy time investment and specialist teams to use. We can now use models “off the shelf” to form predictions of many more formulations in entirely new domains. This pandoras box will not close, and it is clear that statistical model assisted discovery will remain a dominant paradigm in high throughput screening for some time. It is foolish to heed caution in the face of curiosity, and the latest tools will always be misused, misinterpreted, and overgenerously applied by some. This may be to the detriment of the practitioner, and on occasion to the wider

community, but with this field being in its infancy there remain many avenues to experiment and find the correct methods that work best for each individual workflow. If a tool is incorrectly used to discover a material of extraordinary properties then it remains a useful tool. We should always strive for understanding and correctness, but untangling the high dimensional web of decisions that make up the inner workings of a trained neural network into a cohesive whole is akin to boiling the ocean.

The numbers a loss function outputs can be endlessly chased to the sixth decimal place, but these convey little understanding. We must devise our own methods of judging quality. Quality is an elusive property which is difficult to quantify but easier to identify. We believe the metrics presented in this thesis provide a mental avenue for calibrating the outputs of statistical models and information rich materials datasets against human understanding.

The skillsets required in theoretical chemistry and experimental synthesis form a complementary partnership, and it is only by drawing on this union of skills that we gain access to the repository where chemical knowledge is currently stored. It is a simple fact that the professionals who have worked under the constraints of synthetic chemistry for many years have a better understanding of this domain than any AI or numeric metric. Leveraging the knowledge contained within the experts cerebral portfolio remains the only path to truly judging whether a current approach is simply generating numbers which we have decided fit our personal narrative, or is approximating the physical world to such high quality that the approach is worthy of further investment.

References

1. Wei, L. *et al.* TCSP: a Template-Based Crystal Structure Prediction Algorithm for Materials Discovery. *Inorg. Chem.* **61**, 8431–8439 (2022).

2. Song, Y., Siriwardane, E. M. D., Zhao, Y. & Hu, J. Computational Discovery of New 2D Materials Using Deep Learning Generative Models. *ACS Appl. Mater. Interfaces* **13**, 53303–53313 (2021).
3. Malik, S. A., Goodall, R. E. A. & Lee, A. A. Predicting the Outcomes of Material Syntheses with Deep Learning. *Chem. Mater.* **33**, 616–624 (2021).
4. Hu, J. *et al.* MaterialsAtlas.org: a materials informatics web app platform for materials discovery and survey of state-of-the-art. *Npj Comput. Mater.* **8**, 1–12 (2022).
5. Baird, S. G., Diep, T. Q. & Sparks, T. D. DiSCoVeR: a materials discovery screening tool for high performance, unique chemical compositions. *Digit. Discov.* **1**, 226–240 (2022).
6. Zhang, R.-Z., Seth, S. & Cumby, J. Grouped representation of interatomic distances as a similarity measure for crystal structures. *Digit. Discov.* (2022) doi:10.1039/D2DD00054G.
7. Wang, A. Y.-T., Mahmoud, M. S., Czasny, M. & Gurlo, A. CrabNet for Explainable Deep Learning in Materials Science: Bridging the Gap Between Academia and Industry. *Integrating Mater. Manuf. Innov.* **11**, 41–56 (2022).
8. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. Preprint at <https://doi.org/10.48550/arXiv.2112.10752> (2022).
9. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal Diffusion Variational Autoencoder for Periodic Material Generation. Preprint at <https://doi.org/10.48550/arXiv.2110.06197> (2022).
10. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. Preprint at <https://doi.org/10.48550/arXiv.1701.07875> (2017).
11. Alverson, M., Baird, S., Murdock, R. & Sparks, T. Generative adversarial networks and diffusion models in material discovery. <https://chemrxiv.org/engage/chemrxiv/article-details/63694d64fbfd387c25d2d395> (2022) doi:10.26434/chemrxiv-2022-6l4pm.
12. Glawe, H., Sanna, A., Gross, E. K. U. & Marques, M. A. L. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New J. Phys.* **18**, 093011 (2016).
13. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).
14. Li, S. *et al.* Encoding the atomic structure for machine learning in materials science. *WIREs Comput. Mol. Sci.* **12**, e1558 (2022).
15. Mohamad Moosavi, S., Xu, H., Chen, L., I. Cooper, A. & Smit, B. Geometric landscapes for material discovery within energy–structure–function maps. *Chem. Sci.* **11**, 5423–5433 (2020).

16. Hiraoka, Y. *et al.* Hierarchical structures of amorphous solids characterized by persistent homology. *Proc. Natl. Acad. Sci.* **113**, 7035–7040 (2016).
17. Jiang, Y. *et al.* Topological representations of crystalline compounds for the machine-learning prediction of materials properties. *Npj Comput. Mater.* **7**, 1–8 (2021).
18. Bubenik, P. Statistical Topological Data Analysis using Persistence Landscapes. *J. Mach. Learn. Res.* **16**, 77–102 (2015).
19. Adams, H. *et al.* Persistence Images: A Stable Vector Representation of Persistent Homology. *ArXiv150706217 Cs Math Stat* (2016).
20. Gameiro, M., Hiraoka, Y. & Obayashi, I. Continuation of point clouds via persistence diagrams. *Phys. Nonlinear Phenom.* **334**, 118–132 (2016).
21. Smith, P. & Kurlin, V. Families of point sets with identical 1D persistence. Preprint at <https://doi.org/10.48550/arXiv.2202.00577> (2022).
22. Zaki, A., Attia, M., Hegazy, D. & Amin, S. Comprehensive Survey on Dynamic Graph Models. *Int. J. Adv. Comput. Sci. Appl.* **7**, (2016).