Generating new crystals using Diffusion models in the image domain

Tushar Sood

4th Year Project Report
Cognitive Science
School of Informatics
University of Edinburgh

2024

Abstract

Novel functional materials enable fundamental breakthroughs across technological applications from clean energy to information processing. From microchips to batteries and photovoltaics, discovery of inorganic crystals has been bottlenecked by expensive trial-and-error approaches. Recently, computational approaches, and especially deep learning approaches have been leveraged for the task of generating novel crystal structures. Most of these methods have leveraged GNNs or other representation of the material requiring highly custom and intensive algorithm design. This paper seeks to explore the use of image based methods for creating crystals and determining their properties, allowing researchers to leverage the large amount of computer vision algorithm and models that exist to the task of creating novel materials. The results are promising, with diffusion models showing the ability to learn the patterns of atom placement in crystal structures and CNNs showing the ability to learn to predict properties from image representations of crystals. While the time limitations of the project prevent the creation of a novel image diffusion architecture tailored to this task, this study motivates similar works in exploring the potential of image embeddings in the realm of generating novel crystal structures.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Tushar Sood*)

Acknowledgements

I would like to express my deepest gratitude to my supervisor Dr. Daga Panas, as well as Dr. James Cumby and Dr. Sohan Seth for the help, insight and expertise they provided during this project. Without their assistance, this project would not be possible. I would also like to recognise the support of my friends and family in supporting me morally and mentally during the entire research process.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Materials are the cornerstone of most technologies which constitute the modern world and make it possible. For example, graphite serves as the anode material in lithium-ion batteries [1]. Another material, Cadmium Telluride (CdTe) is widely used in thin-film solar cells due to its ability to efficiently absorb sunlight and convert it to electricity [3]. Certain materials, including the aforementioned Cadmium Telluride and Grpahite exhibit highly ordered and symmetrical microscopic structure, which lends them special properties. These materials are called crystals and are the subject of this dissertation and wide ongoing research in Material Science. Crystals exhibit special properties which make many technologies possible, such as transistors, diodes, batteries, solar cells, radiation detectors and communication devices [12]. For these technologies to advance, it is crucial to discover newer crystals which enhance the properties of currently known crystals and which demonstrate new, desired properties. They not only enhance the applicability and power of current technologies, but also pave the way for new technologies.

Traditional approaches for developing new crystals and materials in general have an extremely long time frame, around 10-20 years from initial research to first use [64]. For the last 3 decades, computational approaches such as Computational simulation and Experimental measurements have offered a significant advancement beyond traditional trial-and-error methods in laboratory-based crystal creation processes. However, accelerating materials discovery and design using these methods is challenging due to inherent limitations. Experimental measurement, which involves property and microstructure analysis, is intuitive but time and resource consuming. [64]. Computational simulation, such as electronic structure calculations based on Density functional theory (DFT) is faster but depends on material microstructures, and high-performance computing equipment. In addition, previous calculation results cannot be directly applied when studying a new system. [64]

With the coming of the big data age, there has been an increase in the amount of data available about material structures and properties in the Materials Science community through initiatives such as ICSD [39] [4], Materials Project (MP) [65], OQMD [57],

etc. These databases store computed and measured properties, structures, and other data of currently known crystals. Machine learning, as a tool for finding patterns in high dimensional data is now a promising approach for analysing this data to generate new materials and to predict their properties.

Since the first paper on the application of Artificial Intelligence to the domain of Materials Science in the 1990s, Machine learning has become widely adopted in materials science [64]. Nevertheless, a significant challenge persists: despite the existence of the aforementioned databases, the number of known crystals remains relatively limited. For example, the ICSD which claims to be the world's largest database of completely identified organic structures, only contains approximately 240,000 structures [39]. This is low, particularly concerning the needs of modern machine learning and deep learning approaches, which demand extensive datasets. Moreover, the computational representation of fundamentally infinite crystal structures in a discrete manner remains a significant point of contention. While there are general encoding patterns for crystals, research diverges greatly in the chosen representation methods. Most current representations require novel and highly custom and complicated model architectures to be built. However, major advances in machine learning are often in fields like NLP, Image Processing with GANs, VAEs and diffusion models. Transferring these innovations for use in material generation research often takes years. Hence there is a need to explore representations which can directly use innovations in these fields without having to make heavy modifications to make them usable in materials generation, such as the new and relatively unexplored CrysTens representation [34]. While the image-based approach has previously been utilized for unconditional crystal generation, its potential for property prediction remains to be determined.

In addition to current representations and the required model architectures being complex, the task of directly producing a crystal structure is a very difficult one, as is underscored later, in section 2. A potential and as of yet unexplored method is of indirectly generating a crystal by generating synthetic data such as Electron Charge Density data or X-ray diffusion data which gives direct insights into the underlying crystal structure which makes the synthetic data possible [2]. If we are able to generate synthetic 3D ECD data, it is trivial to then convert that to a crystal struture. Due to the time frame of this project however, it was not possible to work on 3D generation. Working with 2D crystal structures is a step in that direction and can serve to confirm whether models are capable of learning ECD patterns.

1.2 Aims

This project had two aims:

1. Examine the feasibility of image based 2D ECD data generation which can then be used to generate or approximate the underlying crystal structure.
2. Determine if the CrysTens representation is suitable for property prediction. If it is, using classifier guidance, and the existing CrysTens crystal generation framework (Section 3.4), crystals optimised for certain properties can be developed.

1.3 Contributions

I would like to draw the attention of the reader to the following novel contributions presented in the paper:

1. Validated the ability of generative models to learn ECD information from images.
2. Validation of the potency of the CrysTens representation for use in predicting crystal properties.
3. The creating of a dataset of CrysTens representation of stable materials for off-the-shelf use with image generative models.
4. This research expands upon the findings of Alverson et al. [34] and after consultation with Dr. Panas, has been identified as demonstrating publishable potential. The process to prepare this work for publishing is ongoing.

Chapter 2

Background

Improvements in engineering often rely on new materials that suit the problem at hand. From the development of strong, lightweight alloys for airplanes to the creation of biocompatible materials for prosthetics, new materials have pushed the boundaries of what is possible.

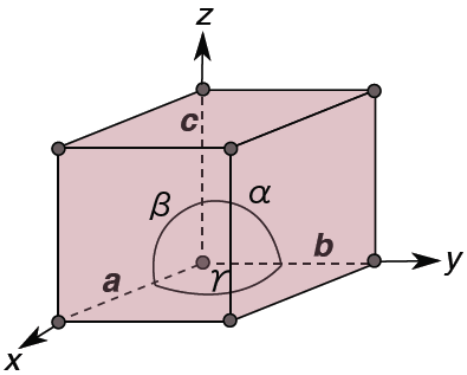
However creating such materials[1] is not an easy task. It requires a deep understanding of the intricate relationships between the properties, functions, and interactions of various elements and structures at the atomic and molecular level.

Traditionally, materials discovery has relied heavily on laborious, time and resource-intensive chemical synthesis in laboratory settings, guided by human intuition and experiments. This process often has large spans between initial experimentation and application of the material, and can be potentially hazardous, involving the use of toxic or volatile chemicals.

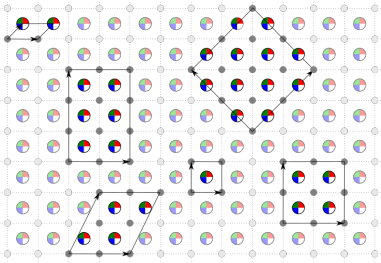
Current computational approaches, which were initially envisioned as an alternative to the traditional lab based research, mostly apply optimisation processes such as random search and bayesian optimization, guided by repeated computations of Density Functional Theory (DFT) to estimate the energy at each iteration of the generation process, searching for the local minima of the energy landscape [64]. However, these repeated DFT calculations are computationally expensive.

By leveraging the power of deep learning algorithms, generative AI can sift through massive datasets of known materials and their properties, and discern patterns and relationships between structures and the properties they have. Using this knowledge, the AI can then propose entirely novel crystal candidates that exhibit tailored properties. This ability to explore a significantly more expansive design space than conventional optimization techniques has the potential to significantly expedite the material discovery process. Furthermore, generative AI can be fine-tuned to optimize materials for specific criteria and constraints. In order to effectively harness this potential, it's important to first understand the foundational principles of crystallography.

[1] "Material" and "crystal" are used interchangeably throughout the report



(a) The 6 lattice parameters. Note that the unit cell edges may or may not be perpendicular to each other. Source: [69]



(b) Different supercells that can be chosen to represent a given structure. Source: [54]

Figure 2.1: Lattice parameters and supercell configurations

2.1 Fundamentals of Crystal structures

2.1.1 Atomic Arrangements

A **crystalline material** is defined as a material in which atoms are arranged in a periodically repeating geometric array, which is called a **crystal structure**. Since these repeating arrangements are far larger than the atoms which constitute them, and these arrangements can theoretically extend infinitely, crystals are often treated as infinite structures, especially since this simplifies calculations of properties [38].

2.1.2 Unit cells

The smallest and most fundamental repeating part of a crystal is the **unit cell**. A crystal is built entirely from repeatedly translating the unit cell along its principle axes, and hence solely determines the symmetry and structure of the crystal. Mathematically, the unit cell has 6 parameters: 3 edge vectors: a, b, c; and the angles between the edges: α, β, γ [5]. This is shown in Figure 2.1a. These vectors and angles can then be used to create fractional coordinates for the points in the unit cell [5].

2.1.3 Planes and properties

Given this definition of crystal structure, crystallographic directions are lines defined using two atoms and a crystallographic planes are planes defined using 3 atoms [38]. Given a crystal, for example NbO2F shown in figure 2.2, it is obvious that some crystallographic planes will have more atoms on them than other planes depending on the points chosen. For example points chosen that lie on the edges will have sparser planes due to the small number of atoms that lie on the edge of NbO2F. For other crystals, this may not be the case. The structure of these planes relates to various physical properties of the crystalline material, such as:

1. Surface Interactions: Adsorption, reactivity, and surface tension depend on the

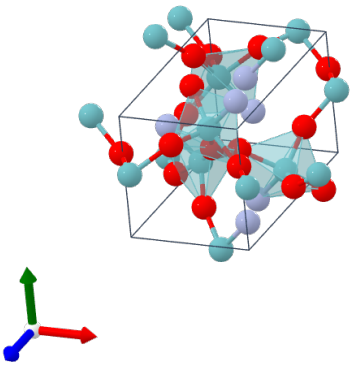


Figure 2.2: Crystal NbO2F, which exhibits higher density at the center and lower atom density at the edges, which means that crystallographic planes constituted primarily by edge atoms contain less atoms than planes made up of primarily center atoms. Source: [65]

availability of atoms at the material's surface, which is higher along dense planes.

2. Defect Behavior: Pores, grain boundaries, crystallite growth, and cleavage planes often align with high-density planes to minimize energy.

3. Plastic Deformation: Crystal deformation occurs more easily along dense planes, as dislocations can move with less disruption to the overall lattice.

2.1.4 Super cells

Supercells are formed by replicating the original unit cell along lattice vectors, typically by integer multiples. They are pivotal in computational simulations for modeling extended systems or exploring defects [36]. For instance, doubling the lattice vectors along each dimension yields an 8 times larger supercell $(2x2x2)$, while preserving the symmetry of the original lattice. There are numerous different supercells that can be chosen for a given material. This is illustrated in Figure 2.1b [36].

2.1.5 Properties

A lot has been said about the properties of crystals, and how the crystal structure is directly liked to the properties a crystalline material has. Table 2.1 lists some material properties and their significance. However, the task of predicting properties of a crystal is difficult because of the following points:

1. The encoding is not sufficient: One of the main difficulties in this research area is the question of how to encode an in principle infinite 3D geometry into a finite, digital representation which despite its finite-ness captures enough features

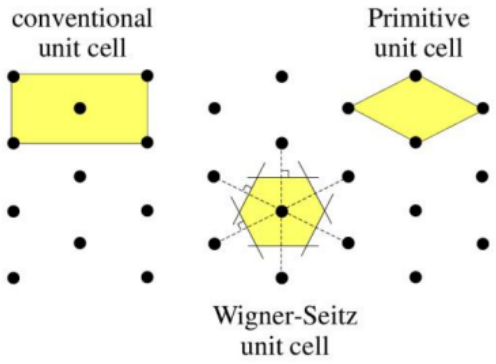


Figure 2.3: Different unit cells for a given crystal structure. Source: [70]

to be of use. Unit cells are often used used for this task, but a unit cell alone may not contain enough information to describe the complex interactions and symmetries of the crystal. For example, in crystals with polar molecules or complex ionic arrangements, long-range interactions between charged units or dipoles can significantly influence properties [58]

2. Many to one function: One crystal structure may have more than one unit cell that can be defined. Different sets of atoms can be chosen to form the unit cell as long as the entire crystal can be rebuilt. Figure 2.3 depicts the different unit cells.

3. Different crystal structures can have similar properties: The many-to-one relationship between crystal structures and material properties poses a challenge for property prediction because it breaks the ideal scenario where each unique structure leads to a unique set of properties [7]. Predictive models also have to search a much wider range of potential structures, increasing computational cost. Additionally, this hints that properties are influenced by subtle factors beyond the crystal structure, complicating the prediction process.

If the task of property prediction is viewed as a function, where the crystal structures are input and the property of interest is the output, the above listed points essentially mean that there is no simple function that can solve this problem. Physically defining a mathematical function to do this is hence infeasible and intractable. Fortunately however, Machine learning models are essentially function approximators, which approximate the function between a pool of input output pairs to maximise the likelihood of the correct output value being produced for an input value.

2.2 Crystal Representation

In order to train machine learning models on the structure to property task, we first need to agree on a salient computational representation of crystal structures which is suitable for machine learning models to use. Since the unit cell might not suffice for representation in the task of property prediction (2.1.5), we should aim to incorporate a comprehensive range of descriptors to provide as much information as we can. The type of descriptors differs throughout literature and there are papers dedicated solely

Table 2.1: Material Properties and Meanings

Property	**Meaning**
Energy above Hull	The energy difference between a material's formation energy and the energy of the most stable reference phase (the hull), indicating its thermodynamic stability relative to other phases. A lower value suggests greater stability [47].
Space Group	Describes the symmetry of the crystal structure, representing the arrangement of atoms or ions within the unit cell. Different space groups exhibit distinct symmetry elements such as rotations, reflections, and translations [50].
Fermi Energy	The Fermi energy, representing the highest energy level occupied by electrons at absolute zero temperature in a material. It is a crucial parameter in determining electronic properties such as conductivity and band structure [51].
Total Magnetic	Refers to the total magnetic moment per unit cell in the material, indicating its magnetic properties. Materials with nonzero total magnetic moments may exhibit ferromagnetic, antiferromagnetic, or other magnetic behaviors [51].
Direct Band Gap	Indicates the direction in which the band gap occurs in the material's electronic band structure. The band gap represents the energy difference between the top of the valence band and the bottom of the conduction band, determining a material's conductivity and optical properties [51].
Indirect Band Gap	Denotes whether the band gap of the material is indirect, meaning that the maximum energy of the valence band and the minimum energy of the conduction band occur at different points in the Brillouin zone. Indirect band gaps typically involve momentum conservation in optical transitions [51].
CBM & VBM	CBM refers to the lowest energy level in the conduction band where free electrons can exist, facilitating electrical conduction. Conversely, VBM is the highest energy level of the valence band, filled with valence electrons under normal conditions. The difference between VBM and CBM is the band gap. CBM and VBM is crucial for designing semiconductors in solar cells and electronic devices, where controlling electron flow and light absorption is essential.[51].

to defining criteria for the descriptor and to analyzing potential descriptors. In deep learning, rather than specifying features explicitly, we allow the model to autonomously learn and extract features directly from the raw data. Despite this, it remains essential to select an appropriate encoding method for representing the crystals. The following are approaches used in literature:

1. SLICES: Simplified Line-Input Crystal-Encoding System (SLICES) encodes the topology and composition of crystal structures into strings, much like how SMILES converts molecular graphs into line notations. More specifically, SLICES

leverages the mathematical concept of ”labeled quotient graphs” to represent periodic crystal structures [31]. The atoms and bonds within a unit cell are mapped to nodes and edges of the quotient graph. Additional labels are assigned to edges indicating the periodic shift vectors required to connect equivalent atoms in neighboring unit cells (Fig. 2.4). However, generating a valid SLICES string requires the model to learn rules unrelated to crystal structure, such as SLICES grammar and atomic order, which adds burden to the generation process, and can result in instability and invalid strings.

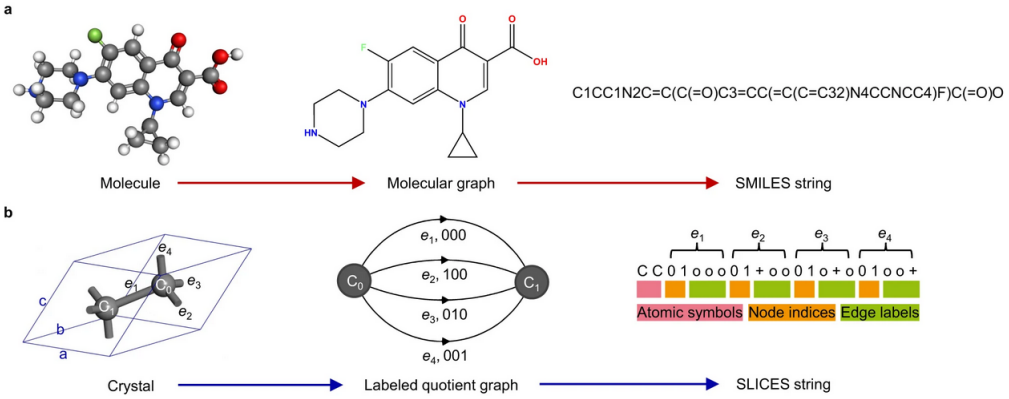


Figure 2.4: Example SLICE string generation [31]

2. Graph Neural Networks (GNN): When crystals are modelled as GNNs, atoms serve as nodes and bonds as edges [56]. Classic architectures like SchNet and Crystal Graph Convolutional Neural Networks (CGCNN) update atom representations based on neighboring atoms and bond lengths, though they have limitations such as overlooking many-body and directional information and having a limited receptive field [45]. Recent developments focus on incorporating geometric details beyond bond length, introducing attention mechanisms, and employing advanced readout functions to enhance information exchange globally [63] [44]. Despite the increasing volume of research and publications on the development and application of GNN models, they continue to be a specialized tool, being relatively challenging to implement, train and transfer to new datasets. Invariance ensures a model's output stays the same even if the input data is slightly modified, such as by rotating or translating the input [55]. Implementing invariance is difficult in graph neural networks since they are not widely supported in ML packages, and are mathematically demanding to implement [20].

3. Indirect generation: However, instead of directly trying to generate the material, we can also try to generate other indicators such as Electron Charge Density (ECD) data (Figure 2.6), from which a crystal structure can be then derived. The charge density of an atom refers to the distribution of electric charge within the atom. In an atom, negatively charged electrons orbit around the positively charged nucleus, which contains protons and neutrons. The charge density describes how the total charge is distributed throughout the atom's volume [2]. When computationally modeled, it often appears as regions of high electron density around the nucleus,

representing where electrons are most likely to be found, and diminishing electron density as you move further away from the nucleus. This distribution resembles a cloud-like structure, with denser regions closer to the nucleus and more diffuse regions further away. Since ECD models the distribution of electrons around an atom, and different atoms have different amounts of electron density and spread, the atom can be inferred [60]. Given an image containing several such blurs, it can be inferred what the underlying atoms are, from which the crystal can be recreated. However, ECD data is limited and 3D, complicating the application of this approach on a full scale. Converting from ECD data to crystals is not difficult, since the same process is how structures of materials are examined in a lab setting, in which X ray crystallography yields ECD data, which is then mapped to a structure [19]. Despite being only a slight modification of standard processes, doing this with synthetic data is as of yet unexplored. Feeding multiple slices from one strucutre into a model to learn is not trivial, and is the ultimate goal in working with this representation.

4. Crystalised Tensors: Another as of yet unexplored representation is images. The CrysTens encoding of a crystal is an image encoding including the pairwise distance matrix, distance graphs, and other chemical information of the crystal, providing a comprehensive and structurally informative data representation (Figure 2.5) [25]. While the pairwise distance graph is not needed for conversion to and from a cif (a file format for crystal structures), this additional information gives models additional guidance to the algorithm and follows the principle of maximising information fed to the model. There are a few challenges with using this representation. Firstly, it's unclear how to effectively incorporate multiple slices from a single crystal structure into a model for comprehensive learning. Additionally, the 32 x 32 x 32 grid format, while unaffected by the number of atoms and rotational symmetries, this might not be the most efficient or informative representation for a machine learning model. As a new encoding, there is only one research paper which employs the CrysTens representation for unconditional generation: Alverson et all [34]. It validates the potential of using this representation for novel crystal generation, and furthermore shows the superior performance of diffusion models over GANs in this task, which suffered from mode collapse, and even with the implementation of Wasserstein loss, were not able to consistently match diffusion models in generating symmetric structures. While crystals generated from synthetic CIF files should be symmetric, due to the stochastic nature of the diffusion process, post-processing was required on the created structures. Post processing sought to average across multiple instances of lattice parameters, atomic assignments and positions to decrease uncertainty. It also employed K means clustering to categorise averaged data and changed atomic number and position values to the nearest centroid to improve output clarity. From this, cifs were generated. These post-processing steps combined with 4 successive relaxations using the Vienna Ab Initio Simulation Package (VASP), produced stable cifs, and overall property distributions which closely matched the distribution of the training data. However, only 6 cif files were selected for further individual assessment, and this low number may not capture the true nature of the individual cif files and may not be a good approximator

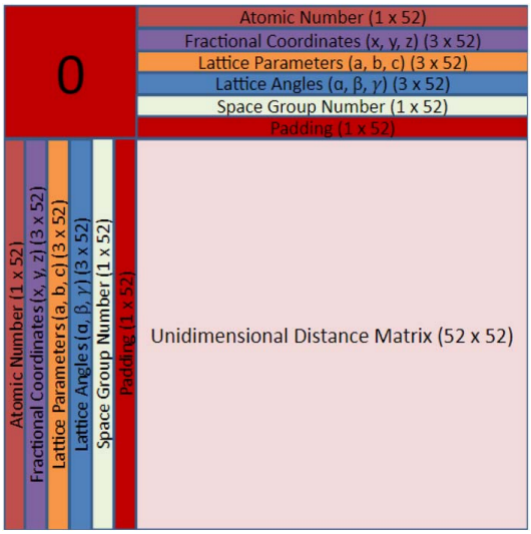


Figure 2.5: The legend for a CrysTens image [34].

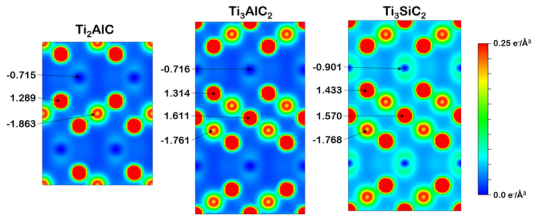


Figure 2.6: An example ECD slice of 3 different crystals [21].

for the performance of the model. Further, more expansive validation should be done. Despite the limited validation, this points towards the use of CrysTens in an image format for the task of inverse design of materials.

2.3 Models used for crystal generation and property prediction

The methods used in Generative AI have been previously used for the reverse task of mapping from the latent space to the crystal itself. The basic flow of all crystal generative models is the same. Crystal structures are input to a deep-learning pipeline after conversion to one of the aforementioned encodings, and then the model learns patterns and relationships within the data to generate novel crystal structures that exhibit desired properties. Material science research has traditionally focused more on the study of molecules than crystals. Due to this many implementations exist for molecules. Since molecules are finite groups of atoms held together by strong chemical bonds, they do not have the same restrictions and considerations, and hence those architectures are not easily transferable to crystals. They will not be discussed when talking about existing literature on the topic.

This section explores the different model architectures that have been used in the task

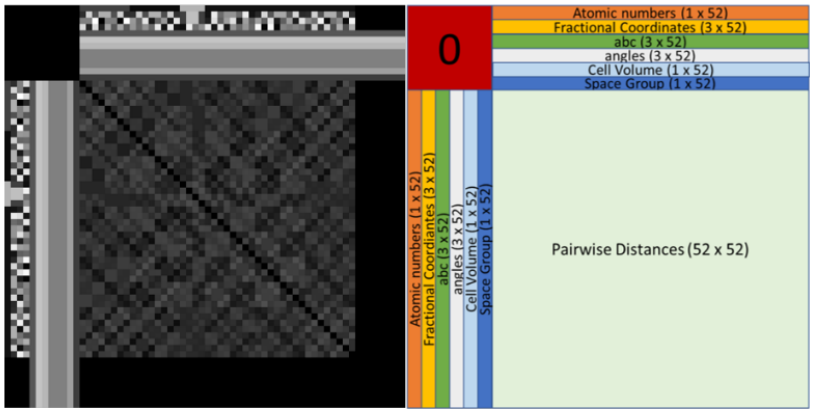


Figure 2.7: CrysTens legend and a sample crystal in CrysTens format

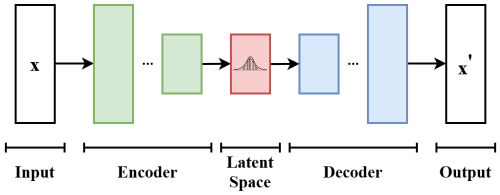


Figure 2.8: VAE Architecture [29]

of crystal generation.

2.3.1 Variational Auto-encoders (VAE)

Variational Auto-encoders (VAEs) are a type of generative model that learn to represent high-dimensional data in a lower-dimensional latent space 2.8. VAEs consist of an encoder network that compresses input data into a latent representation and a decoder network that reconstructs the original data from this latent space [37]. VAEs improve upon traditional auto-encoders because they simultaneously attempt to structure the latent space according to a predefined probability distribution, offering the ability to locate materials with desired properties in the respective regions that the properties inhabit in the latent space and in their intersection [34]. During training, VAEs aim to minimize the reconstruction error while simultaneously regularizing the distribution of the latent space to follow a predefined prior distribution, typically a Gaussian distribution. This regularization encourages the latent space to capture meaningful features of the data and enables the generation of new samples by sampling from the learned latent space distribution [37].

Since there is more research in generating molecules over crystals, there are many more VAE models for generating molecules. Due to this many implementations of VAE exist for molecules [23] [68] [27] [46] [61]. The one key VAE model which exists for generating new crystals using VAE approach is a Crystal Diffusion Variational Auto Encoder (CDVAE), which uses a Diffusion model (Section 2.3.3) as the decoder in the VAE network, and encodes materials as grpah neural networks (GNN). The CDVAE

framework integrates an encoder, property predictor, and decoder, trained concurrently. The encoder, a SE(3) equivariant periodic graph neural network (PGNN), compresses materials into a lower-dimensional latent space, from which the property predictor forecasts key material properties [24]. The decoder, a noise conditional score network diffusion model, reconstructs stable structures from perturbed inputs, handling noise in atom types and coordinates. Utilizing equivariant diffusion models as decoders enables direct manipulation of atomic positions, making CDVAE adaptable across diverse chemical elements and structures. Post-training, new materials are generated by sampling the latent space and gradually denoising randomly placed atoms to match the training data distribution, focusing on stable materials to enhance stability [24]. A multi-layer perceptron (MLP) takes as input the latent representation and predicts the following 3 properties: density, energy, and number of unique elements [24].

One weakness of the paper is that while the paper addresses scale in-variance it does not touch upon shift in variance and how that is achieved in the model architecture. Additionally, the properties that are predicted are not as extensive as the properties that are required in industrial or research applications, as suggested by the list of properties available in Matbench [43], a leader board for comparing property prediction results from algorithms.

Despite this, when introduced, CDVAEs outperformed the existing models, such as 3D Voxel based models, wherein materials are treated as 3D voxel images, and vector based approaches where atom coordinates, types and lattice are encoded as vectors. 3D Voxel based models result in low validity and these models are not rotationally invariant (meaning rotating a 3D Voxel causes the model to think it is an entirely different Voxel). The latter models are generally not invariant to any Euclidean transformations [26].

2.3.2 Generative Adversarial Networks (GAN)

Generative Adversarial Networks (GANs) employ a competitive framework between two neural networks: the generator and the discriminator. The generator learns to synthesize realistic data samples from noise, while the discriminator attempts to distinguish these from real examples. This adversarial training drives both networks to improve, with the goal of the generator ultimately producing highly plausible synthetic data [53]. A key advantage of GANs is their ability to implicitly learn the data's probability distribution, rather than relying on a predefined model as in VAEs [49]. Hence GANs excel in open-ended generation tasks where defining an explicit probability distribution is difficult. However, GANs can be notoriously challenging to train, often experiencing instability and convergence issues. Additionally, they may suffer from mode collapse, where the generator only produces a limited range of samples, failing to capture the full diversity of the data distribution.

Nouira et al. introduced Crystal-GAN as the first GAN application in material science. Crystal GAN is a Cycle GAN based approach for generating crystal structures [9]. Cycle GANs consist of **two** generator and discriminator networks, where the generators aim to map images from one domain to another while the discriminators attempt to distinguish between real and fake images [8]. The key benefit of a Cycle GAN in this task is its ability to learn the mapping between two domains without paired image

training data. Cross domain learning refers to producing synthetic data that is different from the training data, such as predicting three-element crystals from two-element training data. Nouira et al. aimed to use this model to produce novel ternary metal hybrids (A–B–H phases) from observed A–H and B–H binary structures. Using lattice geometry and ionic atom positions as input data, the model first generates mixed domain pseudo-binary samples [9].

Previous models like the one by Nouira et al. focused on general inverse design principles, but recent advancements extend cross domain learning [67]. However, the ultimate goal of inverse design is to directly discover materials with desired properties. While early encodings relied solely on composition or implicit chemical properties, recent research shows that incorporating physical properties like band gap and formation energy as input conditions is essential for true inverse design. Dan et al. developed a Wasserstein Generative Adversarial Network (WGAN), dubbed the 'MatGAN model', to predict chemically valid hypothetical materials using input data with specific properties [35]. The model, consisting of discriminator and generator deep neural networks (DNNs), successfully predicted two million materials, with 92.53% being novel. Despite limitations in material representation, model performance was assessed using various methods, achieving 84.5% compliance with basic chemical rules. Long et al. introduced a constrained-crystal deep convolutional GAN (CCDCGAN), integrating a constraint network to optimize the generator and predictor without embedding material properties in the input [35]. The constraint network ensures that the generated materials adhere to certain constraints essential for the desired output, such as structural stability, chemical validity, or other physical properties. The CCDCGAN outperformed traditional GANs in generating stable structures. Additionally, Zhao et al. developed a DNN-based Generative Model (GM), CubicGAN, trained on cubic space group structures to simplify model design. The model used chemical properties, coordinates, and space groups as input data, and discovered 506 crystal structures [18].

All the GAN models mentioned so far aim to minimise the Wasserstein distance, which measures the dissimilarity between the distribution difference of the predicted and real materials. Models trained on specific crystal structures (e.g., cubic space groups) have limited applicability to broader material design. Due to challenges in cross-domain generalization, a separate model might be required for each distinct class of materials. In order for these models to generalise and to be used on a broader subset of the material space, and for more feasible structures to be predicted, physics-guided principles and constraints must be incorporated in addition to the Wasserstein distance. Physics Guided crystal structure generation model (PGCGM) aims to address this by integrating affine matrix, space group symmetry information, and an atomic distance matrix/loss calculation module to ensure structural accuracy and symmetry [33]. The physics guided loss calculations maintain accuracy of physical and chemical properties by controlling atomic distances to prevent unfeasible structures. This model demonstrates high structural diversity and symmetry, with 1869 out of 2000 materials successfully optimized.

2.3.3 Diffusion Models

Diffusion models are a powerful class of generative models which learn the dynamics of a diffusion process to generate data. The diffusion process (also known as the forward process) progressively degrades an input sample x_0 via transition kernels $q(x_t|x_{t-1})$, defining a Markov chain $x_0 \to x_1 \to \ldots \to x_T$, where $T \in N$ is the number of diffusion steps and $1 \leq t \leq T$ [6].

The use of diffusion models in materials science has expanded significantly since their introduction, surpassing the earlier popularity of GANs. Despite this, as a recent addition, there is limited research in this field [34] [15]. The aforementioned Crystal Diffusion Variational Auto-encoder (CDVAE) is the first example of Diffusion models applied to this problem, and uses diffusion model as the decoder which produces crystal structures. In this approach, the components of a crystal structure unit cell (number of atoms, lattice parameters, and composition) are diffused independently, without explicit collaboration or information sharing. Since these are actually intricately linked, independent diffusion ignores this relationship, and potentially hinders the models ability to generate physically realistic and stable structures. Separate models for diffusion also increases computational cost. Models like MatterGen improve upon earlier approaches by introducing joint diffusion and significantly expanding optimizable properties. This includes all scalar properties, symmetry, and composition, a substantial leap from CDVAE's focus on density, composition, and number of atoms [15].

A pure diffusion model for generating new crystal structures is UniMat [32]. UniMat condenses atoms in a material's unit cell by storing their continuous x, y, z coordinates at corresponding entries in the periodic table. It represents crystals in a 4-dimensional space [L, H, W, C], where L = 9 and H = 18 represent periods and groups in the periodic table, and C = 3 denotes the x, y, z locations of each atom. Using this representation, a diffusion model can effectively move atoms from random initializations to target locations in a unit cell. Despite lacking explicit structure modeling, UniMat generates high-fidelity crystal structures from larger, more complex chemical systems, surpassing previous graph-based approaches across various generative modeling metrics. Despite its superior performance compared to CDVAE, as shown in Figure 2.9 from the paper, one limitation of the UniMat model is that since it uses the periodic table for representing materials, in smaller chemical systems the representations will be sparse, which might result in increased computational requirements.

Comparative investigations between diffusion models and Generative Adversarial Networks (GANs) indicates that diffusion models consistently produced more stable structures, particularly when stability is a key objective, while offering better control over the generation process [34] [1, 4] Furthermore, their design inherently reduces the issue of "mode collapse" observed in GANs, leading to a more diverse exploration of the solution space [34].

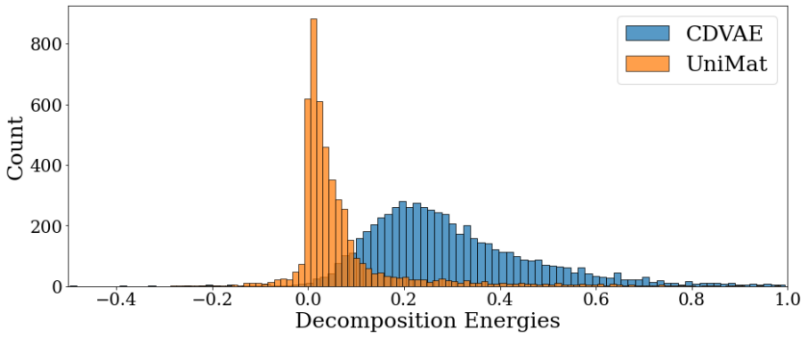


Figure 2.9: The distribution of formation energies of materials generated by UniMat was consistently lower than the materials generated by CDVAE. Lower decomposition energy indicates greater thermodynamic stability. [32]

Chapter 3

Methodology

3.1 Combined Data Strategy

This project's data processing can be divided into two phases: combined and separate. During the combined phase, both research arcs shared a similar data processing approach focused on retrieving, filtering, and storing data. In the separate phase, the processes diverged, where research arc 1 involved retrieving and creating ECD images, while research arc 2 focused on generating xtal2png images. The combined data processing strategy is outlined below, while the specific separate strategies are detailed in their respective sections.

Obtaining data was one of the most challenging aspects of this project. Firstly, While there exist many databases for crystals, such as: Materials Project, Alexandria, OQMD, ICSD, COD and AFLOW, giving the illusion of data abundance, this is not true. Firstly, there is a significant overlap between each of these databases, where multiple databases have the same crystal. Moreover, most of these databases except the Materials Project, do not contain ECD data that is central to the first research arc [1] of this project. Normally lack of data can be addressed by data augmentation approaches, especially for images. However, since the CrysTens images model highly ordered arrangements of atoms and contain contextual features such as distance matrices and unit cell volume, which may get distorted, Data Augmentation was not used. This was resolved by an advanced data augmentation strategy used later in the process (Sec. 3.3.1).

Secondly, the databases often have different methods of calculating properties. Research projects with more resources, such as MatterGen, recompute properties through DFT calculations following Materials Project guidelines for all entries in datasets to ensure identical data consistency. Due to limited resources, this was not possible for this project.

Converting materials to the CrysTens representation for the second research arc [2], as

[1] First research arc: Explore the potency of image based ECD data generation which can then be used to generate or approximate the underlying crystal structure.

[2] Second research arc: Determine if the CrysTens representation is suitable for property prediction. If it is, using classifier guidance with the existing CrysTens crystal generation framework by Alverson et al.,

well as the ECD data augmentation process for the first research arc, were tedious and time consuming processes. While the resulting datasets do not have large memory footprints, the generation process is time consuming.

3.2 Database Selection

There are 6 big crystal databases: Materials Project, Alexandria, OQMD, ICSD, COD and AFLOW that are available for use. From these, the ICSD is a closed source commercial database which requires a license, so this is automatically excluded. COD is an open database, based on user submission, which raises concerns about the quality, accuracy and completeness of the data in it. Dealing with these concerns complicates data handling and hence COD was not used. AFLOW's specialized data access methods, requiring custom libraries and potentially complex stability calculations, made it less suitable for this project's focus on rapid data integration from multiple sources.

In the end the following databases where used:

- MP (v2022.10.28, Creative Commons Attribution 4.0 International License), an open-access resource containing DFT-relaxed crystal structures obtained from a variety of sources, but largely based upon experimentally-known crystals [17] [48] [66] [52]. The Materials Project hosts an S3 Bucket on AWS under the AWS Open Data Sponsorship program which contains ECD data. This dataset was used to derive the ECD dataset used in research arc 1, and the Materials project in general also contributed some structures for research arc 2.
- The Alexandria dataset (Creative Commons Attribution 4.0 International License), an open-access resource containing DFT-relaxed crystal structures from a variety of sources, including a large quantity of hypothetical crystal structures generated by ML methods or other algorithmic means [16] [30] [28] [22]. This database contributed data towards research arc 2.

Crystal structures were retrieved from the Materials Project using the legacy MP-API along with their existing calculations of properties. This process resulted in a combined dataset of 155361 unique structures. From here, the processing of the data differed between each research arc.

3.3 Research Arc 1: ECD

We mentioned that data for the ECD was fetched through the MP-API. For each object which had ECD data, this returned a Chgcar object, a file format for storing electron density information. While all real charge distributions are made up of discrete charged particles, they are often approximated as continuous scalar functions of position for the convenience of modeling [2]. This is represented in Pymatgen [3] as a layer of slices which contain the ECD data at that slice in the object. This is essentially just a way

crystals optimised for certain properties can be developed.

[3]Pymatgen is a Python library designed for materials analysis, providing tools to work with crystal structures, molecules, electronic structure calculations, and more.

Number of Unets	1
Dimension Multiplication factors	1,2,4
Attention Layers	False False True
Cross Attention Layers	False True True
Channels	1
Batch size	4
Image size	770 x 770
Number of training steps	100000

Table 3.1: Google Imagen parameters

of making continuous data discrete, and is akin to sampling a function at continuous intervals.

Preliminary inspection on this data yielded that the ECD of whole crystals was often sparse, meaning numerous slices were blank. These blank slices correspond to no electrons at these locations and provide important 3D information about how the charge density is diffused around the atoms and the space between the atoms. However, if we use slices individually, not as part of a 3D structure, they are not useful since the large amount of black slices contributes negatively to the model. Due to this, these sparse slices were excluded for the unconditional model which generates one slices individually, leaving 14447 slices.

These 14447 slices are standardized for the pixel values to be between 0 and 1, and the resulting image is resized to 256x256 and converted to gray-scale, and then used to train a Google Imagen model with the parameters described in 3.1. The Google Imagen model used was the unofficial implementation from lucidrains on Github [4].

Due to constraints on computational resources, only one UNet was chosen. Cascading UNets, where one UNet with creates an image with smaller dimensions and the second UNet increases the resolution of the image, might give better results [62].

3.3.1 Data Augmentation for single slice generation

The main problem encountered in the previous approach was a lack of data. 14447 slices is not sufficient to train a large model such as Google Imagen. A novel data augmentation strategy tailored to ECD data was developed and used.

From the 155361 unique structures, that we had at the end of the combined data processing phase, all unstable structures (energy above hull greater than .1) were filtered out leaving only stable structures. For the ECD Data, prior to data augmentation, ECD data was fetched through the API for all stable materials with less than 52 sites. The novel data augmentation strategy only required crystal structures to generate ECD image slices. The structures were only chosen from the Materials Project, since each structure generates $\frac{n!}{(n-3)!}$ slices, where n is the number of atoms int he unit cell of the structure. If n is 20 (the mean number of atoms in the unit cell of structures in the Materials Project), this produces 6840 slices. Due to the large yield of slices per

[4] Available at https://github.com/lucidrains/imagen-pytorch

structure, only using 10% of the MP dataset with random selection yielded sufficient ECD slices. Hence, there was no need to use structures from the Alexanderia dataset.

As mentioned previously, X ray diffusion is used to measure electron charge density in a material. From the locations of electron density, the positions of atoms are deduced. The data augmentation strategy reversed this process. There might be a lack of ECD data, but there is not a lack of crystal structures. If the process of converting from ECD data to crystal structure is akin to denoising to locate the position of atoms, the reverse process is akin to simply adding noise. "Noise" was added by defining a Gaussian distribution, a "glow" extending from the position of atoms. The exact data augmentation procedure is as follows.

1. Load the structure of a material.
2. Calculate the average volume per atom. Average volume per atom = cell volume/number of sites.
3. Scale the unit cell edges by volume/atom to give a standard volume/atom of 1 $Ang^3/atom$. Unnormalized crystal structures exhibit non-uniform atomic clustering, with some regions densely packed and others sparsely populated. This inconsistency necessitates volume scaling of the unit cell prior to enlargement to ensure accurate analysis and modeling.
4. Pick three random atoms in the structure and use them to define a plane. This is a crystallographic plane as described in 2.1.3. The density of the atoms in this plane can be used to predict physical properties of the structure along this plane.
5. Convert the structure into a supercell big enough to be larger than the plane. To create a supercell, the unit cell is replicated in the x, y, and z dimensions 3 times. For improved computational efficiency, the algorithm selects plane-defining points from the unit cell rather than the supercell. This minimizes redundant calculations, as the supercell's larger size increases the likelihood of multiple point combinations yielding the same plane.
6. Find all atoms in the supercell that are within a tolerance of .1 Ang. from the plane.
7. Project these atoms onto the plane. The plane was trimmed to size 40 Å x 40 Å.
8. Apply a uniform gaussian blur around atom positions. Ideally, the spread of the blur would match the electron charge density of the atom we model, but keeping the spread of the blur constant simplified the project and made it more fitting for the project timeline.

Steps 3,5 and 7 were done using the Pymatgen package in python.

3.3.1.1 Results Comparison

The following metrics were used to analyse resulting images from our models. Since the images generated represent 2D slices, we are unable to verify if it composes a sensible

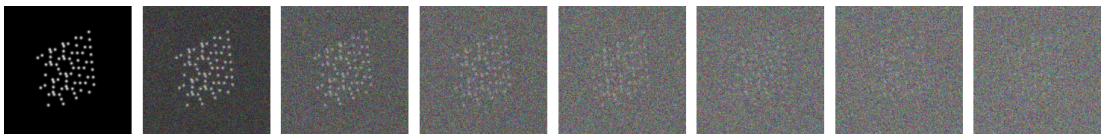


Figure 3.1: The forward diffusion process, wherein noise is added until the image resembles pure noise. The reverse diffusion process learns to denoise the image back. This is a random image yielded from the data augmentation strategy.

structure, but the hope is that the model is able to learn what slices of plausible and possible structures look like and generate those.:

1. Kolmogorov–Smirnov test: The Kolmogorov-Smirnov test is a non-parametric statistical method for comparing distributions [10]. It evaluates the maximum difference between the empirical distribution of a sample and a known reference distribution. Since the data generated across models might have different distributions, which do not all meet the requirements of parametric tests, the KS Test is a robust metric to use as it makes minimal assumptions about the distributions under consideration.

2. Fréchet Inception Distance (FID): The Fréchet Inception Distance (FID) is a tool to measure how realistic images generated by . It does this by comparing the features extracted from both a set of real images and a set of model-generated images, using a pre-trained image recognition system like Inception V3. If the features are very similar, the FID score will be low, and a higher FID score signified a bigger difference between generated and real image features, suggesting the images are different.

To evaluate the generated images, the distribution of the number of atoms in each image was compared between the generated and training data using the Kolmogorov–Smirnov test and the generated images were compared with the training images using the FID score. Since it is not feasible to count the number of atoms in each generated image ourselves, the python OpenCV library was used to create a script which automatically did this. First the grayscale images were deblurred to reduce noise, then thresholding was done to isolate features, which are identified as contours. These contours are then counted, which is the number of atoms present in the image. Primary focus with regards to interpretation of model results was put on visual inspection as FID scores might not be most relevant given the specialised nature of the data, and the fact that the Inception model is pre-trained on images of a vastly different domain (ImageNet), meaning that compared to the dataset that the Inception model is trained on, our dataset lacks diversity.

3.4 Research Arc 2: CrystTens pure property prediction

The demand is not just for new materials, but for those with specific, user-defined properties relevant to applications like semiconductors, solar cells, and batteries. The CrysTens representation has shown promise in unconditionally generating materials that resemble the training dataset. The next step is to assess if CrysTens can predict

material properties. If successful, a classifier could guide image generation, enabling conditional material creation. This task, termed 'Pure Property Prediction,' focuses solely on property prediction, as opposed to 'Integrated Property Prediction' which aims to generate novel materials with desired characteristics. Pure property prediction is a pre-requisite to Integrated property prediction.

CrysTens is a way to encode crystal structures as images. The CrysTens representation itself takes the form of a 64x64x4 tensor. The topmost 12 rows and the leftmost 12 columns encode structural information extracted directly from the CIFs. This includes the atomic numbers for each unique atom present in the crystal (listed up to 52 atoms), their fractional coordinates (x, y, z), lattice parameters (a, b, c), lattice angles (α, β, γ), and the space group number [25]. If a crystal has fewer than 52 unique atoms, the remaining space in this section is filled with zeros.

The first layer encodes pairwise Euclidean distances between each atom using a dedicated 52x52 matrix. Subsequent layers focusing on each dimension separately. Here, distance graphs depict the relative uni-dimensional positions of all atoms. For the remaining three layers, distance graphs are employed for each dimension, illustrating the relative distances between atoms along individual axes. While not strictly necessary for CIF to CrysTens conversion, as only the top most 12 layers are required for conversion, providing additional information enhances the model's understanding and can can help guide the model in its learning process.

The rationale behind CrysTens' design emphasizes a combination of both structural and interatomic information. The structural component ensures that symmetry, atomic basis information, and lattice parameters are all encoded into the representation. The interatomic component ensures that the spatial relationships between atoms are captured. This dual approach aims to make it easier for the generative model to learn how to create new, realistic crystals that adhere to both fundamental structural principles and plausible atomic arrangements.

A few final points about the design are worth noting. The choice of a 64x64 layer dimension is primarily due to common deep learning heuristics favoring powers of two; optimization of this size is a potential area for future research. The current implementation of CrysTens can only accommodate crystals with 52 unique atoms or less.

This following was the research process for the research on Property Prediction with Crystallised Tensors representation:

3.4.1 Data

Filtering using the MP API with large datasets is not recommended, so all filtering was done locally after the structures were retreived. From the 155361 materials, only materials with less than 52 sites were kept for the task of crystal structure property prediction since a 64x64 CrysTens image can only model structures with 52 sites. This resulted in 53786 structures being discarded, leaving 101575 structures. These structures were used for the task of energy above hull prediction. However, for the prediction of all other properties, all materials with energy above hull less than .1

eV/atom were dropped. Predicting properties for stable materials only allowed us to reduce the spread of the data, decreasing problem difficulty and increasing performance. Since unstable materials have limited practical use, this decision was compatible with the project's goals. This resulted in a further 17669 materials being dropped, leaving 83906 materials for the task of property prediction (all properties except energy above hull prediction). These materials were then converted to CrysTens representation using the xtal2png package in Python.

In contrast to Materials Project, which was queried via the MP-API, the Alexandria database was fully downloaded onto the local machine. The database is itself fairly compact, at 1.926 GB (the JSON files, not the final images), as compared to other datasets used for machine learning, such as ImageNet, which is 150 GB [42]. The download was a collection of JSON files. Using python scripts, each JSON file was first examined for defects such as missing data. Then, chemical data was extracted from each file. This was done in batches so as to not cause memory overload by storing so much information in python variables. Each file contained around 100000 structures, and these were saved in the original CIF format in pickled arrays, one by one. Then, each pickled array (which corresponds to one JSON file), was unpickled and each structure was converted to CrysTens format using the Xtal2png package in python. Each CrysTens image was stored in a folder by itself. Next, all the files were moved to a central folder, and as they were moved, the material ID was matched to the pickled array, and the filename and corresponding properties were saved into a CSV. Leveraging a CSV file to track each structure's properties and corresponding filename streamlined the data handling process. We simply provided the CSV and the data folder containing all slices to the training script. By controlling arguments within the training script, we could dynamically select specific data subsets, such as stable materials only. This eliminated the need for creating redundant and space-consuming datasets with significant overlap. For training to predict energy above hull, the datasets from MP and Alexanderia were merged together and the model was trained to predict the energy above hull for all structures, regardless of stability. For other properties, only stable structures were used, as mentioned in the previous section. This process was very time-consuming, due to the size of the dataset and the one-by-one processing it required. For example, iterating through all JSON files and extracting material data and storing it in pickled arrays took 5 hours, and the process of un-pickling and then creating the CrysTens images took around 12 hours. The subsequent moving images to a central location while performing lookup to match filename and properties took approximately 15 hours [5].

The most crucial property for material prediction is stability, measured by energy above hull. This quantifies the likelihood of a structure decomposing, with low energy above hull indicating stability. Consequently, models for properties other than stability are trained exclusively on stable structures. The Filtered-Alexanderia and Filtered-MP datasets were combined based on energy above hull.

Preliminary Data Analysis yields that numerous properties such as energy above hull, number of sites and band gap are skewed. Transforming and standardising was explored

[5]Times are on AMD Ryzen 7 4700U (2.00 GHz) with 16 GB RAM

on these properties through a number of experiments done to ascertain whether this led to an increase in performance. It was hypothesised that this decreases the variance in the data, the effect of outliers and makes the distributions more normal. The results of these experiments are given in the Results section. Energy above hull was not scaled (as it may be negative in some generated crystals) and not transformed either because it naturally has a smaller variance. Furthermore, since the raw distribution does not resemble a poisson or logarithmic distributions, it is not appropriate to do these transforms on it. Due to this, raw energy above hull was used.

When actually training the models, a data split of 64%-16%-20% was used.

3.4.2 Baseline model training

Next, a basic baseline model was designed and implemented in Keras, and used to predict all properties. A diagram of the model is available in the Appendix.

This model uses 2D convolutional layers with ReLu activations for feature extraction, followed by max pooling for down-sampling. Dense layers further process these features. The final output layer employs a single neuron with a linear activation function for the regression task. Small 3x3 kernels allow for an enlarged receptive field with fewer parameters compared to larger kernels, enhancing computational efficiency. MAPE was chosen as the target metric due to its differentiability, and ability to be compared across datasets (scaled/not scaled, transformed/not transformed), since its units are percentages. MSE, MAE and RMSE were also recorded.

3.4.3 Pre trained models

Pretrained models offer a promising avenue for feature extraction in machine learning, leading to improved performance and reduced training times in areas with data scarcity. It might be possible to extend this to materials science. Transfer learning, a prevalent technique in this domain, involves fine-tuning models initially trained on different datasets for specific tasks down the stream. This approach not only addresses data scarcity but also reduces the computational burden compared to training models from scratch. Interestingly, even pre-trained models at intermediate training stages can outperform fully trained counterparts when utilized as feature extractors [41].

An initial investigation employing VGG16 from Keras revealed a significant computational burden, requiring approximately 50 minutes per training epoch. Due to the project's limitations in computing resources, comprehensive training and evaluation of VGG16 became infeasible. Additionally, training results obtained with more complex models suggested that VGG16, boasting 138.4 million parameters, might be overly intricate for the specific feature extraction task at hand. Consequently, a more lightweight model, MobileNetV2 with only 3.5 million parameters, was chosen for further exploration. The workflow used was:

1. Instantiate base model with loaded ImageNet weights.
2. Freeze all trainable layers

3. Use this model for feature extraction and build a smaller model to use these features as input

It was hypothesised that this would yield better results as a result of more thorough feature extraction. The model used was derived from the baseline model but with a few changes. First, all convolutional layers in the baseline models were replaced by the MobileNetV2 from Keras. Second, more and larger dense layers were added to accommodate the increase in convolutional power, and to allow the dense layers to understand the output of the convolutional layers. In order to mitigate the risk of overfitting for the now significantly larger model, dropout was added between the dense layers.

3.4.3.1 Model Performance

The following metrics were used to analyse model performance:

1. MSE: MSE measures the average of the squared differences between the true and predicted values. It penalize large errors more than small errors and is hence sensitive to outliers. Furthermore, it is also not in the same units as the target variable.
2. MAE: MAE measures the average of the absolute differences between the true and predicted values. It provides a more intuitive measure of error as it is not affected by the direction of the errors. Less sensitive to outliers compared to MSE.
3. RMSE: RMSE is the square root of the MSE, providing a measure of the average magnitude of the error in the predicted values. It is in the same unit as the target variable, making it more interpretable. Since RMSE penalises large errors more than small ones, it is sensitive to outliers.
4. MAPE: Mean Absolute Percentage Error measures the average magnitude of the errors between predicted values and actual values as a percentage of the actual values. This metric expresses error as a percentage, making it easy to interpret and particularly useful for comparing the accuracy of prediction models across different data scales. Since MAPE weighs all errors in proportion to the true values, it is especially sensitive to cases where the actual value is near zero, potentially leading to disproportionately high errors.

Chapter 4

Results

4.0.1 ECD

In contrast to the next section, due to the more straightforward aims of this section combined with the enormous computational cost to repeatedly train a Google Imagen model, the model was trained once using hyperparameters listed in table 3.1 for the original basic sliced data, and once with the same hyperparameters for the data yielded from the data augmentation strategy.

Due to the small amount of training data in the original dataset, the large Imagen model overfit to the data (Fig 4.1), as well as having subpar image quality. The poor performance of this approach was apparent in the visual inspection and did not necessitate further numerical analysis.

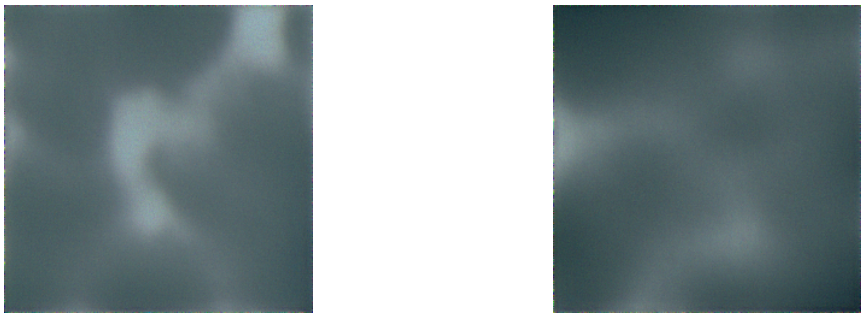
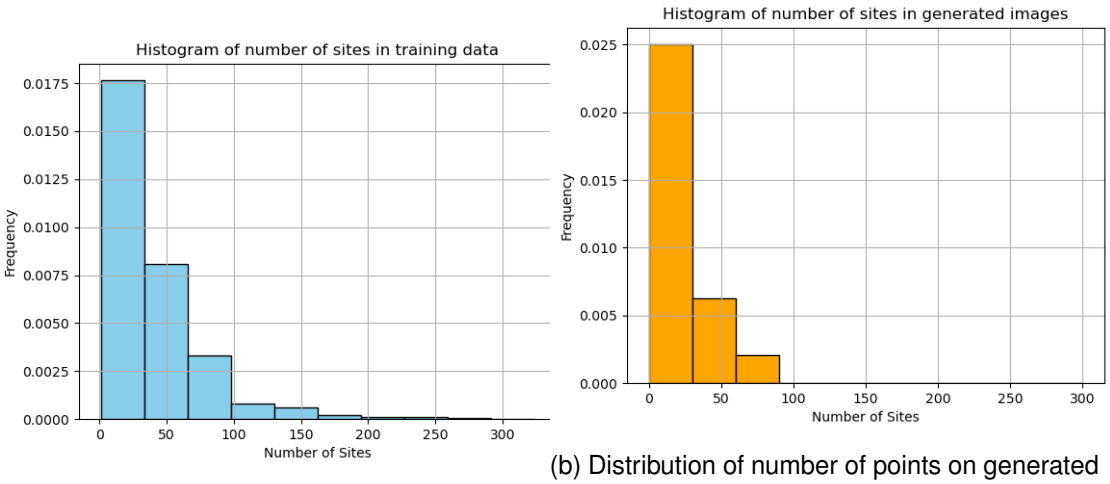


Figure 4.1: 2 images with poor quality sampled from the Imagen model at the end of the training process on the raw ECD data.

It was to ameliorate this problem that the novel ECD data augmentation strategy was used. The results of the Image generation process are shown below, with more samples in appendix x.

This model trained on the data augmentation strategy yields much better image results based on visual inspection. Once the model was trained for 100,000 training steps, 1000 images were sampled from the model and analysis done on each image to identify the number of atoms present in the image. In order to compare the distribution on the number of atoms, a KS test with an alpha of .001 was done. The KS test yields a p-value



(a) Distribution of training number of points

(b) Distribution of number of points on generated slices

Figure 4.2: Comparison of number of points per slice

of .02 and a KS statistic of .45, which signifies similarity between the two distributions. In statistics, the value of alpha chosen is somewhat subjective to the problem and domain. Since the KS test in effect ”simulates” distributions by approximating an EDF to them, distributions with more points will be smoother and hence more closely approximate the distribution, whereas with fewer points the EDF will be coarser. This discrepancy can make it difficult to compare the two distributions, especially if the true underlying distributions are very similar but not identical, as in our case. Hence, a lower alpha was chosen to introduce leniency into the evaluation of the results. Following up the KS test with a visual analysis, the distribution of number of sites in the generated slices is ”pinched”, and has a smaller variance than the training distribution, as shown in Figure 4.2. This is most likely because the model does not have enough samples to learn the data points at the edges of the distribution. Nevertheless, the distributions appear highly similar from a visual analysis. For visual comparison, some images from the training and generated sets are given here. Visual analysis of the images shows that the model learned the patterns in atom distribution well, and the generated images closely resemble the training images. While the generated images are similar to the training set, they exhibit less distinct, highly geometric patterns (Fig 4.3) which exist in the training images. The less frequent appearance of these patterns might be due to the difficulty in learning the complex structure and patterns embedded within these representations. Additionally, there are less of such images in the training data, which might cause data scarcity in attempting to create these images. The FID score of the two sets of datasets was extremely large, at 10023.6. This is most likely due to the concerns mentioned in the Methodology, such as lack of diversity, lack of transferability of the FID score to such a unique and specific image dataset, and the difference in the ImageNet deatset on which the InceptionV3 model is trained and our generated images.

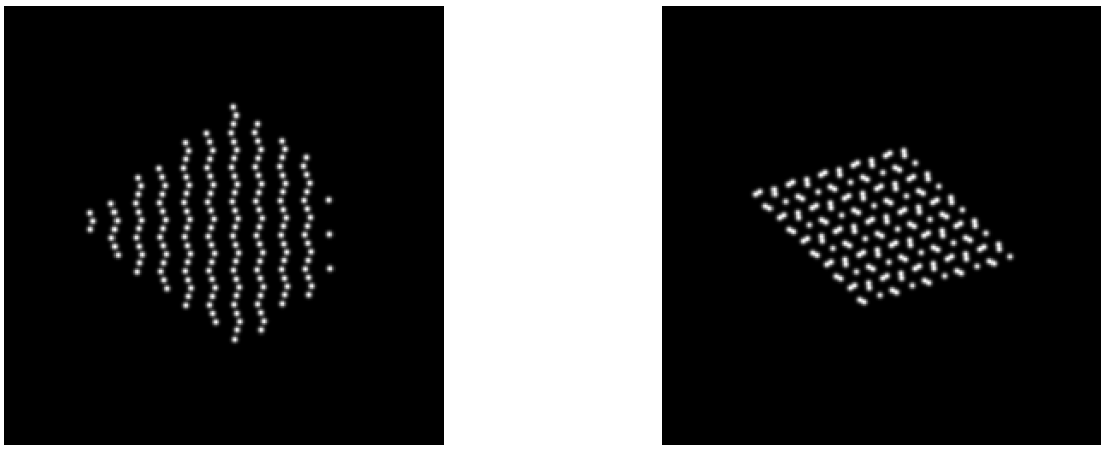


Figure 4.3: 2 examples of the ”highly geometric” ECD slice images from the training data

4.1 Property Prediction

The following are the results for the baseline model for predicting properties (Table 4.1).

Table 4.1: Summary of Properties and Metrics (Rounded)

Property	MAPE	Test MAE	St dev of predictor	Variance
Energy above hull	132.885	0.023	0.031	0.001
Density	8.853	1.132	2.799	7.833
CBM	93.666	1.397	1.922	3.693
VBM	99.890	1.379	2.262	5.117
Fermi energy	95.802	1.723	2.650	7.022
Energy per atom	28.036	4.089	8.032	64.510
Indirect band gap	94.398	0.341	0.868	0.753
Direct band gap	80.897	0.415	0.883	0.780
Density of states	2641.707	3.729	6.778	45.945
Total energy	31.238	14.551	43.356	1879.734
Total magnetisation	10437.273	1.667	3.967	15.737

These results are non-trivial, indicating that the CrysTens representation encodes enough information to predict material properties. For example, Pandey et al. achieve an error rate of .05 eV/atom using a graph neural net approach to predict energy above hull, as compared to .023 eV/atom for our model [13]. Olsthoorn et al. achieve an MAE of .388 eV and MAPE of 8.4% when predicting band gap of materials from the Organic Materials Database as compared to MAE of .415 and .341, and MAPE of 94% and 80% for Direct and indirect bandgap from our baseline model [11]. Comparison yields that our model performed well on these properties, but encountered difficulties when predicting electronic properties. For example Venturi et al. achieve a MAE of .193 eV and .180 eV for CBM and VBM prediction respectively using a stacked model approach as compared to 1.397 and 1.379 for our baseline model [14]. Fermi energy

demonstrates a similar high MAPE and MAE. It is not prudent to directly compare our values to the values yielded from literature, as most studies use different datasets for final testing. Furthermore, for most properties, our experiments were done only on stable materials. In literature, some authors have this restriction and some do not. While we cannot directly compare results, by looking at the difference in the results, we can estimate the difference in power of the approaches. If we were to include unstable materials, the variance for band gap would only increase by approximately 3 eV, from approximately 17 eV to 20 eV. The subsequent increase in data points (approximately 1.5 million additional points) might even improve performance. The critical obstacle in this project was computing resources. Effectively tripling the amount of data would lead to an unsustainable increase in computing cost. Hence this was not explored.

However, this research arc focuses on the feasibility of property prediction with this specific representation, not in generating SOTA results. In this regard, the research demonstrates the potential for success since a simple, ”toy” baseline model, which can be implemented relatively quickly compared to the GNNs used in other papers gives some results that are comparable to literature. It was hypothesised at this point that using a pre-trained model as a feature extractor would greatly increase performance. Furthermore we utilised the same model architecture for all properties, where as the papers that we compare our results to are focused solely on predicting one property. Implementing more complicated models which incorporate the latest advances in Computer Vision with additional data as well as tailoring the architecture to this task specifically has the potential of giving much better results. Overall, while the results are encouraging, they also suggest that refinements are needed to reduce the errors, as many results deviate by roughly one standard deviation, and have relatively high MAPEs.

4.1.1 Standardisation and Transformation model results

Interestingly, although it was originally expected that properties with higher variance would exhibit a higher Mean Absolute Percentage Error (MAPE), the results are mixed. Out of the four variables which have higher variances compared to other variables (energy per atom, density of states, total energy and total magnetisation), only two follow the hypothesis (Density of states and total magnetisation). Total energy, with a sizable skew and variance, has comparatively good MAPE values for the baseline model. In properties with high variance, gradient descent’s sensitivity to scale can amplify this issue, especially when dealing with high-variance or large-scale target variables, potentially leading to unstable training and vanishing/exploding gradients. It was hypothesised that standardising these target variables to decrease range, and a square root transform to make target variables normal. Experiments were done across a subset of properties to evaluate this hypothesis. The properties were chosen to maximise the diversity of distributions and to decrease the number of properties that needed to be evaluated. 3 runs for each property were deemed salient in order to maximise the accuracy of results, and keep computational load low. The results confirmed the hypothesis, and are displayed in Table 4.2.

Only MAPE was used to compare between unstandardized/untransformed results and standardised/transformed results. The results confirmed the hypothesis that transforming

and standardising would help variables with high skew and large range. Direct Band Gap, a distribution with heavy skew and low range, was impacted positively by the addition of transforming, but was impacted negatively by standardisation with an increase in MAPE. This is expected as the primary issue in the distribution of band gap is the skew, not the spread. Density of states on the other hand showed a decrease in MAPE upon standardising, but a substantial increase upon transforming. This contradicts the hypothesis as Density of states exhibits a large skew and a large variance, and it was expected that standardising and transforming would yield the best results. Likewise, Total magnetisation, a distribution very similar to density of states, showed best performance when only standardized. The extremely high MAPE values might be linked to the fact that when the actual values are very small, the MAPE can be arbitrarily high.

In accordance with these results, only these properties from the Alexandria dataset with extreme range were modified: indirect and direct band gap were square root transformed, density of states was square root transformed and standardized, and total magnetisation was solely standardized.

Table 4.2: MAPE Values for Different Properties and Transforms

Property	**Transform**	**MAPE**
band_gap_dir	None	80.8974
	SQRT only	52.7551
	STAND only	104.3302
	SQRT and STAND	104.1719
dos_ef	None	2641.707
	SQRT only	1285.0458
	STAND only	105.293
	SQRT and STAND	100.4395
total_mag	None	10437.2725
	SQRT only	6682.8726
	STAND only	95.6248
	SQRT and STAND	126.0211

Table 4.3: The effect of standardising and transforming on certain predictors.

4.1.2 Pre-trained model results

Next experiments were conducted to evaluate the hypothesis that using pretrained models as feature extractors by freezing the weights would lead to a decrease in error.

In addition to MAPE, the MSE, MAE and RMSE of all predictors are given in table 4.5.

The results indicate that using a larger feature extraction pipeline did not benefit the regression process proportional to the increase in parameters and training time.

The mean change in errors is -1%, 10% and 34% for the MAE, MSE and RMSE respectively, and 28% for the MAPE, while accounting for a 10 fold increase in the

Table 4.4: MAPE and Percentage Change from Baseline for Various Properties. For properties that were standardised/transformed, the comparison is done with the transformed/standardised value given in the last section.

Property	**Pre-trained Model MAPE**	**% Change from Baseline**
Energy above hull	153679.200	115548.090%
Indirect band gap	43.546	-53.870%
Direct band gap	44.076	-45.520%
Density of states	1266.272	-52.070%
Total energy	31.200	-0.120%
Total magnetisation		-100.000%
CBM	232.464	148.180%
VBM	148.701	48.870%
Density	33.326	276.420%
Fermi energy	132.455	38.260%
Energy per atom	33.726	20.0359%

Table 4.5: Percentage Changes in MAE, MSE and RMSE from the baseline model for the pretrained model trained for 5 epochs

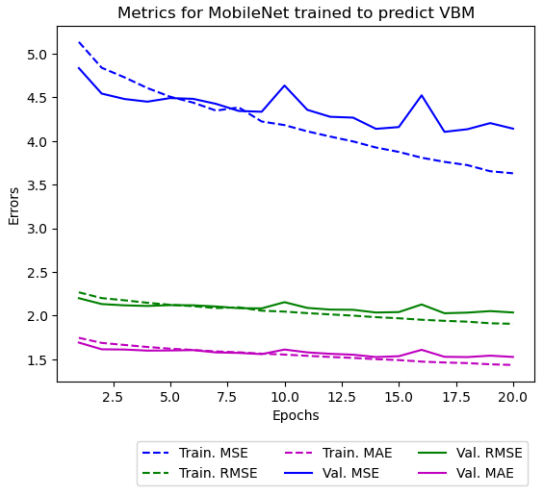
Property	Percent change in MAE	Percent change in MSE	Percent change in RMSE
e_above_hull	4%	-3%	-2%
band_gap_dir	86%	98%	88%
dos_ef	10%	20%	9%
energy_total	-1%	-10%	-298%
total_mag	-4%	-9%	-4%
cbm	-21%	-43%	-20%
vbm	-45%	-123%	-40%
density	-3%	-9%	-74%
efermi	-5%	-5%	-2%
e_per_atom	-44%	29%	-79%

amount of time taken to achieve this performance. The increase in time is due to the increase in trainable parameters from the bigger dense layers, since experimentation showed that the complexity of the dense layers must be increased in order to match the increase in the convolutional layers. The model shows largest increases of errors in predicting density and CBM. Analysing the baseline results, it is clear that the model consistently struggles with predicting CBM and VBM, demonstrating a much higher MAE than the literature. CBM and VBM are influenced by specific chemical bonding and electronic states near the Fermi level, which may not be fully captured by the features encoded by the CrysTens representation. While the features that are part of the CrysTens representation provide sufficient structural information (which corresponds to low error in band gap, density and energy per atom predictions), they lack electronic structure which would be greatly beneficial in predicting electronic properties such as CBM, VBM and Fermi energy. This is also another potential cause for the poor performance in predicting CBM, VBM and Fermi energy. Since Band gap is the difference in CBM and VBM (meaning they are closely related), a multi-modal approach could be used to first predict band gap, since the model performs better in predicting band gap, and then the band gap could be supplied as a feature in addition to the CrysTens representation to predict the CBM and VBM. This highlights that CrysTens needs to be bolstered when predicting electronic properties, which is identified as a weakness of the representation.

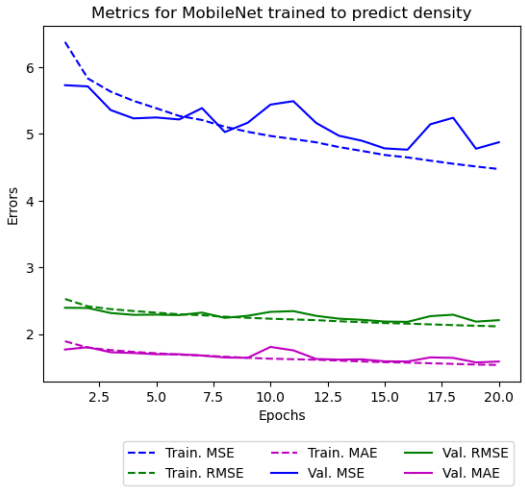
The exceptional increase in MAPE for energy above hull is likely because of the small actual values being a divisor in the calculation, especially since the MAE, MSE and RMSE did not exhibit such a large increase. Both types of band gap show an decrease in error. The model shows marginal increases or decreases for other properties, with the exception of energy per atom where there is a drastic reduction in the error. The performance decrease might be a result of undertraining; larger models with more parameters likely need longer training periods to fully train all parameters. While the model training metrics for some properties exhibited validation metrics being higher than training metrics towards the last epoch, which initially was thought to be suggestive of overfitting, validation loss can sometimes be higher than the training loss temporarily for one or two epochs. Hence, increasing the number of epochs might lead to improved performance. However, the project's computational resources restricted our ability to conduct such extended training.

To evaluate the possibility of this claim, the pre-trained model's performance on predicting CBM, VBM from the MP dataset for stable materials was evaluated for 20 epochs, as opposed to 10 epochs (with an early stopping condition of 5 epochs on validation loss). The results are shown in Table 4.6. The results indicate that while training for longer epochs does indeed decrease error. But again, the decrease in error is not justified by the increase in training time for training 15 additional epochs. Additionally, the training plots show the validation loss rising above the training loss and the difference between them steadily increasing thereafter (Figure 4.4). The validation loss plateaus with minimal improvement, suggesting possible over-fitting. In summary, the pre-trained models yield lackluster results.

Usually institutions involved in material science AI which undertake such research have access to high compute power. While their access allows them to sustain extensive



(a) The training metrics for the pre-trained MobileNet model learning to predict VBM from the MP dataset



(b) The validation metrics for the pre-trained MobileNet model learning to predict density from the MP dataset

Figure 4.4: Performance of MobileNet during extended evaluation of 20 epochs on property prediction

training times, this presents a significant barrier to entry for smaller research groups.

Table 4.6: Change from Pretrained Model at Different Epochs

	% Change from 5 Epochs of pre-trained model			% Change from orig. baseline results		
Property	MAE	MSE	RMSE	MAE	MSE	RMSE
CBM	1%	0%	0%	4%	9%	4%
VBM	-7%	-13%	-7%	12%	25%	12%
Density	-3%	-2%	-1%	41%	120%	39%
Band gap	-3%	3%	2%	-5%	-4%	-1%

Chapter 5

Discussion

5.1 Conclusion

We have shown that machine learning generative models have the capacity to learn the position of atoms in crystallographic planes in 2D. Our research also shows that it is feasible to predict properties from Xtal2Png images, although the prediction of electronic properties might required tailored architectures or multimodal approaches. While our research did not demonstrate exceptional results in property prediction, some results are comparable to existing methods with a simple CNN model. More complicated models are likely to yield even better results.

5.2 Future Works

5.2.1 ECD

The present findings from Electron Charge Density (ECD) modeling highlight the promising trajectory of diffusion models in producing innovative ECD datasets. A clear avenue for enhancing ECD generation lies in the adoption of a 3D diffusion model for data fitting. Expanding the modeling framework to encompass three-dimensional ECD data would markedly enrich the information accessible to researchers, as a single slice per material may prove inadequate for effectively leveraging the generated ECD data for meaningful insights or applications. Hence, transitioning towards a more comprehensive, three-dimensional approach holds significant potential for advancing our comprehension and utilization of ECD data for material generation. While this greatly en

5.2.2 CrysTens property prediction

The current property prediction process can be improved by incorporating chemical features into the CrysTens images. Existing methods often use this approach to add additional chemical features not easily and already incorporated in the representation [59]. by incorporating more information, we can provide a richer material representation

for the model to learn from. However, for Machine learning approaches to be relevant in property prediction, the cost of generating new chemical descriptors on datasets and their incorporation with CrysTens images must be smaller than using traditional DFT-related means to predict properties [40]. Larger architectures coupled with Data Augmentation from other restricted databases such as the ICSD should also be examined. Larger models with more parameters, such as models with attention mechanisms and transformers, have greater ability to capture complex patterns and dependencies in the data. However this also means that these models have a larger number of parameters to fit which requires more data to train them.

Bibliography

[1] Graphite Anode Materials: Natural & Artificial Graphite. URL https://www.targray.com/li-ion-battery/anode-materials/graphite.

[2] *Essential Principles of Physics*. John Murray, 1978. ISBN 9780719533822.

[3] Introduction, 2011. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527633708.ch1. Section: 1 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527633708.ch1.

[4] Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, 2013. ISSN 1543-1851. doi: 10.1007/s11837-013-0755-4. URL https://doi.org/10.1007/s11837-013-0755-4.

[5] *Solid State Chemistry and its Applications*. John Wiley & Sons, 2nd edition, 2014. ISBN 978-1-119-94294-8.

[6] Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015. doi: 10.48550/arXiv.1503.03585. URL https://doi.org/10.48550/arXiv.1503.03585.

[7] *Crystal Science Fundamentals*. Springer, 2017. doi: 10.1007/978-94-024-1117-1_1.

[8] Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. URL https://arxiv.org/abs/1703.10593.

[9] Crystalgan: Learning to discover crystallographic structures with generative adversarial networks. *Preprint*, 2018. URL https://arxiv.org/abs/1810.11203.

[10] Small-sample corrections to kolmogorov-smirnov test statistic. *Pioneer Journal of Theoretical and Applied Statistics*, 15:15–23, 2018.

[11] Band gap prediction for large organic crystal structures with machine learning. *Advanced Quantum Technologies*, 2(7-8):1900023, jul 2019. doi: 10.1002/qute.201900023. URL https://doi.org/10.1002/qute.201900023.

[12] Things You Don't Know about Crystals: Application of Crystalline Material-NATURAL SCIENCES:Taiwan Research Highlight, 2019. URL https://trh.gase.most.ntnu.edu.tw/en/article/content/10.

[13] A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):123, 2020. doi: 10.1038/s41524-020-00362-0. URL https://www.nature.com/articles/s41524-020-00362-0.

[14] Machine learning enabled discovery of application dependent design principles for two-dimensional materials. *Machine Learning: Science and Technology*, 1(3), 2020.

[15] Analysis of the weighted kappa and its maximum with Markov moves. *arXiv preprint arXiv:2010.00232*, 2020. doi: 10.48550/arXiv.2010.00232.

[16] Crystal graph attention networks for the prediction of stable materials. *Sci. Adv.*, 7:eabi7948, 2021. doi: 10.1126/sciadv.abi7948. URL https://doi.org/10.1126/sciadv.abi7948.

[17] A framework for quantifying uncertainty in dft energy corrections, 2021. Working Paper, Version 1.

[18] High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 2021. doi: 10.1002/advs.202100566. URL https://doi.org/10.1002/advs.202100566.

[19] X-ray crystallography, August 28 2022. URL https://chem.libretexts.org/@go/page/55906.

[20] Graph neural networks for materials science and chemistry. *Commun Mater*, 3:93, 2022. doi: 10.1038/s43246-022-00315-6. URL https://doi.org/10.1038/s43246-022-00315-6.

[21] Interfacial stabilities, electronic properties and interfacial fracture mechanism of 6h-sic reinforced copper matrix studied by the first principles method. *Crystals*, 12(1):51, 2022. doi: 10.3390/cryst12010051. URL https://doi.org/10.3390/cryst12010051. Received: 29 November 2021; Revised: 25 December 2021; Accepted: 27 December 2021; Published: 30 December 2021.

[22] A dataset of 175k stable and metastable materials calculated with the pbesol and scan functionals. *Scientific Data*, 9:64, 2022. doi: 10.1038/s41597-022-01177-w. URL https://doi.org/10.1038/s41597-022-01177-w.

[23] Improving vae based molecular representations for compound property prediction. *J Cheminform*, 14:69, 2022. doi: 10.1186/s13321-022-00648-x. URL https://doi.org/10.1186/s13321-022-00648-x.

[24] Crystal diffusion variational autoencoder for periodic material generation. *arXiv*, 2110.06197, 2022. doi: 10.48550/arXiv.2110.06197. URL https://doi.org/10.48550/arXiv.2110.06197. Accepted to ICLR 2022. Code and data are publicly available at this https URL.

[25] xtal2png documentation, 2022. URL https://xtal2png.readthedocs.io/en/latest/. Accessed on March 2 2024.

[26] Crystal diffusion variational autoencoder for periodic material generation. Technical report, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA, 2023. Available from: txie@csail.mit.edu, xiangfu@csail.mit.edu, oct@csail.mit.edu, regina@csail.mit.edu, tommi@csail.mit.edu.

[27] RGCVAE: Relational graph conditioned variational autoencoder for molecule design. *arXiv*, 2305.11699, 2023. doi: 10.48550/arXiv.2305.11699. URL https://doi.org/10.48550/arXiv.2305.11699. Published as arXiv:2305.11699 [cs.LG].

[28] Machine-learning-assisted determination of the global zero-temperature phase diagram of materials. *Advanced Materials*, 35(23):210788, 2023. doi: 10.1002/adma.202210788. URL https://doi.org/10.1002/adma.202210788.

[29] Vae basic — wikimedia commons. https://en.wikipedia.org/wiki/Variational_autoencoder#/media/File:VAE_Basic.png, 2023. Accessed: 10-April-2023.

[30] Symmetry-based computational search for novel binary and ternary 2d materials. *2D Materials*, 10(3):035007, 2023. doi: 10.1088/2053-1583/accc43. URL https://doi.org/10.1088/2053-1583/accc43.

[31] An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, 14(1):7027, 2023.

[32] Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235*, 2023. doi: 10.48550/arXiv.2311.09235.

[33] Physics guided deep learning for generative design of crystal materials with symmetry constraints. *npj Comput Mater*, 9:38, 2023. doi: 10.1038/s41524-023-00987-9. URL https://doi.org/10.1038/s41524-023-00987-9.

[34] Michael Alverson, Sterling Baird, Ryan Murdock, and Taylor Sparks. Generative adversarial networks and diffusion models in material discovery, November 2022.

[35] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. URL https://arxiv.org/abs/1701.07875.

[36] H. Arnold. *Transformations of the Coordinate System (Unit-Cell Transformations)*. International Union of Crystallography, 2006. doi: 10.1107/97809553602060000510.

[37] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antônio Barros da Silva, and Sérgio Lima Netto. *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer Nature, 2021. ISBN 3030706796, 9783030706791.

[38] Robert Vincent Coleman and K. Lark-Horovitz, editors. *Solid state physics*, volume 11. Academic Press, New York, 1974.

[39] Zagorac D, Muller H., S. Ruehl, J. Zagorac, and S. Rehme. Recent developments in the inorganic crystal structure database: Theoretical crystal structure data and related features, 2019. URL https://journals.iucr.org/j/issues/2019/05/00/in5024/.

[40] James Damewood, Jessica Karaguesian, Jaclyn R. Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. Representations of materials for machine learning, January 2023.

[41] Andong Deng, Xingjian Li, Di Hu, Tianyang Wang, Haoyi Xiong, and Chengzhong Xu. Towards inadequately pre-trained models in transfer learning. Accepted by ICCV'2023, 2023. URL https://arxiv.org/abs/2203.04668.

[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[43] Alexander Dunn, Qi Wang, Alex Ganose, Daniela Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6:138, 2020. doi: 10.1038/s41524-020-00406-3. URL https://doi.org/10.1038/s41524-020-00406-3.

[44] Sheng Gong, Tian Xie, Yang Shao-Horn, Rafael Gomez-Bombarelli, and Jeffrey C. Grossman. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *arXiv*, 2208.05039, 2022. doi: 10.48550/arXiv.2208.05039. URL https://arxiv.org/abs/2208.05039. [Submitted on 9 Aug 2022 (v1), last revised 27 Mar 2023 (this version, v3)].

[45] Sheng Gong et al. Examining graph neural networks for crystal structures: Limitations and opportunities for capturing periodicity. *Sci. Adv.*, 9:eadi3245, 2023. doi: 10.1126/sciadv.adi3245.

[46] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv*, 1610.02415, 2016. doi: 10.48550/arXiv.1610.02415. URL https://doi.org/10.48550/arXiv.1610.02415. Published as arXiv:1610.02415 [cs.LG].

[47] Geoffroy Hautier. Data mining approaches to high-throughput crystal structure and compound prediction. In Seda Atahan-Evrenk and Alan Aspuru-Guzik, editors, *Prediction and Calculation of Crystal Structures*, Topics in Current Chemistry. Springer, Cham, 2013. doi: 10.1007/128_2013_486.

[48] Matthew Kristofer Horton, Joseph Harold Montoya, Miao Liu, and Kristin Aslaug Persson. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. *npj Computational Materials*, 2019.

[49] Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *Journal Name*, 23(116):1–43, 2022.

[50] International Union of Crystallography. *International Tables for Crystallography*, volume Volume A. International Union of Crystallography ; Springer, 1st online edition, 2006. URL http://link.springer.com/10.1107/97809553602060000100.

[51] International Union of Crystallography. *International Tables for Crystallography*. International Union of Crystallography ; Wiley, 2nd online edition, 2013. ISBN 978-1-118-76229-5. doi: 10.1107/97809553602060000113. Volume D: Physical properties of crystals.

[52] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1): 011002, 2013. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

[53] Mlaan Jovanović and Mark Campbell. Generative artificial intelligence: Trends and prospects. *IEEE Computer*, 55(10):107–112, Oct. 2022. doi: 10.1109/MC.2022.3192720.

[54] KasugaHuang. 2d supercell example, 2023. URL https://commons.wikimedia.org/wiki/File:2d_supercell_example.svg. [Online; accessed April 17, 2024].

[55] Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. *arXiv*, 1905.04943, 2019. doi: 10.48550/arXiv.1905.04943. URL https://arxiv.org/abs/1905.04943. [Submitted on 13 May 2019 (v1), last revised 24 Oct 2019 (this version, v2)].

[56] Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(18), 2024. doi: 10.1186/s40537-023-00876-4. URL https://doi.org/10.1186/s40537-023-00876-4.

[57] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):1–15, December 2015. ISSN 2057-3960. doi: 10.1038/npjcompumats.2015.10. URL https://www.nature.com/articles/npjcompumats201510. Publisher: Nature Publishing Group.

[58] Charles Kittel and Paul McEuen. *Introduction to Solid State Physics*. John Wiley & Sons, 2018.

[59] Chao Liang, Yilimiranmu Rouzhahong, Caiyuan Ye, Chong Li, Biao Wang, and

Huashan Li. Material symmetry recognition and property prediction accomplished by crystal capsule representation. *Nature Communications*, 14, 2023. doi: 10.1038/s41467-023-40756-2. URL https://doi.org/10.1038/s41467-023-40756-2. Published online: 25 August 2023.

[60] LibreTexts. Electronic structure of atoms, January 15 2024. URL https://chem.libretexts.org/@go/page/469204.

[61] Jaechang Lim, Seongok Ryu, Jin Wook Kim, et al. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform*, 10:31, 2018. doi: 10.1186/s13321-018-0286-7. URL https://doi.org/10.1186/s13321-018-0286-7.

[62] Hongying Liu, Xiongjie Shen, Fanhua Shang, and Fei Wang. Cu-net: Cascaded u-net with loss weighted sampling for brain tumor segmentation. *arXiv preprint arXiv:1907.07677*, 2019. doi: 10.48550/arXiv.1907.07677. URL https://arxiv.org/abs/1907.07677. 9 pages, 4 figures.

[63] Yichao Liu, Zongru Shao, and Nico Hoffmann. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv*, 2112.05561, 2021. doi: 10.48550/arXiv.2112.05561. URL https://arxiv.org/abs/2112.05561.

[64] Yue Liu, Tianlu Zhao, Wangwei Ju, and Siqi Shi. Materials discovery and design using machine learning. *Journal of Materiomics*, 3(3):159–177, September 2017. ISSN 2352-8478. doi: 10.1016/j.jmat.2017.08.002. URL https://www.sciencedirect.com/science/article/pii/S2352847817300515.

[65] Materials Project. Materials project homepage, 2023. URL https://next-gen.materialsproject.org/.

[66] Jason M. Munro, Katherine Latimer, Matthew K. Horton, Shyam Dwaraknath, and Kristin A. Persson. An improved symmetry-based approach to reciprocal space path selection in band structure calculations. *npj Computational Materials*, 2020.

[67] Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chemical Science*, 19, 2020.

[68] Tomoki Ochiai, Toshihiro Inukai, Masayuki Akiyama, et al. Variational autoencoder-based chemical latent space for large molecular structures with 3d complexity. *Commun Chem*, 6:249, 2023. doi: 10.1038/s42004-023-01054-6. URL https://doi.org/10.1038/s42004-023-01054-6.

[69] University of Cambridge. Crystallography - parameters, 2024. URL https://www.doitpoms.ac.uk/tlplib/crystallography3/parameters.php.

[70] University of Illinois. Physics 460 lecture 4, 2016. URL https://courses.physics.illinois.edu/phys460/fa2016/Physics%20460_lecture4.pdf.

Chapter 6

Appendix A

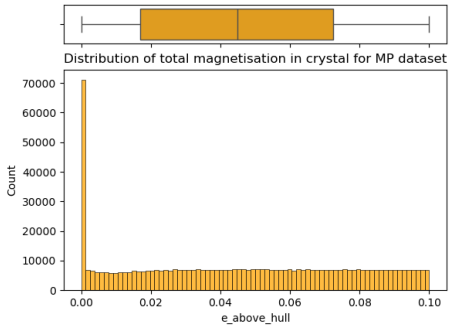
6.1 Preliminary Data Analysis

Table 6.1: Statistics of raw Alexandria data

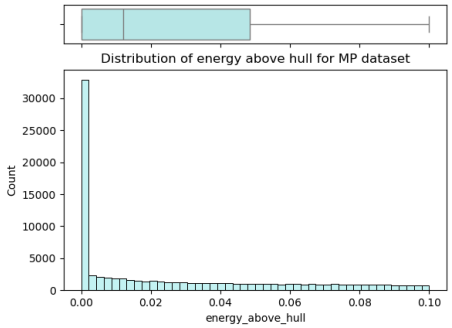
Properties	E. above hull	Ind. Band gap	Dir. Band Gap	DOS EF	Energy Tot.	Tot. Mag.	# sites
Mean	0.0451	0.2983	0.4315	6.4074	-48.62	1.4104	9.4896
Median	0.0450	0.0260	0.1342	4.4554	-36.95	0.0000	8
St dev	0.0312	0.8677	0.8829	6.7783	43.3560	4.4480	6.5084
Variance	0.0010	0.7529	0.7795	45.9453	1879.7385	19.7849	42.3594
Range	0.1000	17.7628	18.3179	308.7864	532.7524	196.7300	51
Skewness	0.0157	0.9413	1.0102	0.8639	-0.8080	0.9513	0.6866
# of Outliers	0	86951	77618	39332	33686	97372	21343
% of Outliers	0.0000	1.5102	1.3481	0.6831	0.5851	1.6912	0.3707

Table 6.2: Raw summary statistics of Materials Project data

	E above hull	**band gap**	**density**	**cbm**	**vbm**	**efermi**	**energy per atom**
Mean	0.0263	1.0970	5.7973	3.9564	2.0697	3.4923	-9.7246
Median	0.0120	0.0697	5.1428	3.9446	2.0363	3.4323	-7.2981
St dev	0.0307	1.5376	2.7987	1.9218	2.2621	2.6498	8.0318
Variance	0.0009	2.3642	7.8328	3.6933	5.1172	7.0216	64.5103
Range	0.1000	17.8914	24.2425	20.0971	28.9530	31.8895	86.3262
skewness	1.3897	2.0044	0.7015	0.0184	0.0443	0.0680	-0.9063
# of Outliers	0	2560	1882	926	1142	401	13910
% of Outliers	0	3.0510	2.2430	1.1036	1.3610	0.4779	16.5781

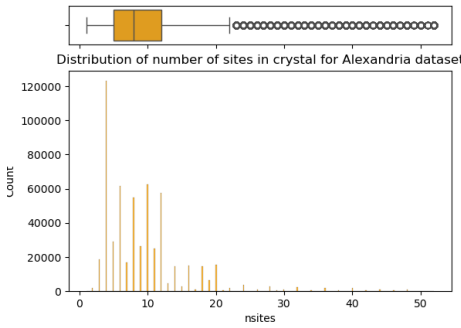


(a) Distribution of Energy above hull in the Alexanderia dataset

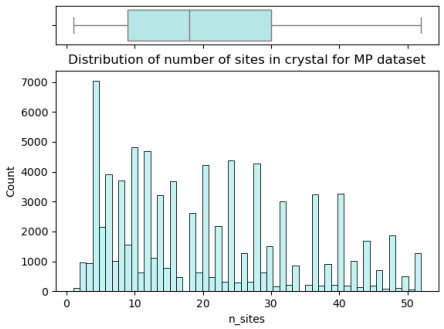


(b) Distribution of Energy above hull in the MP dataset

Figure 6.1: Energy Above Hull

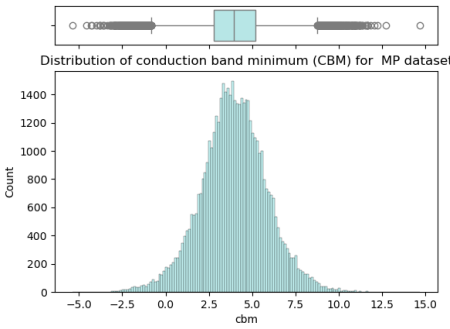


(a) Distribution of number of sites in the Alexanderia dataset

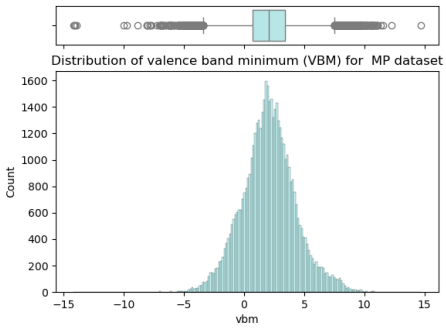


(b) Distribution of number of sites in the MP dataset

Figure 6.2: Number of Sites

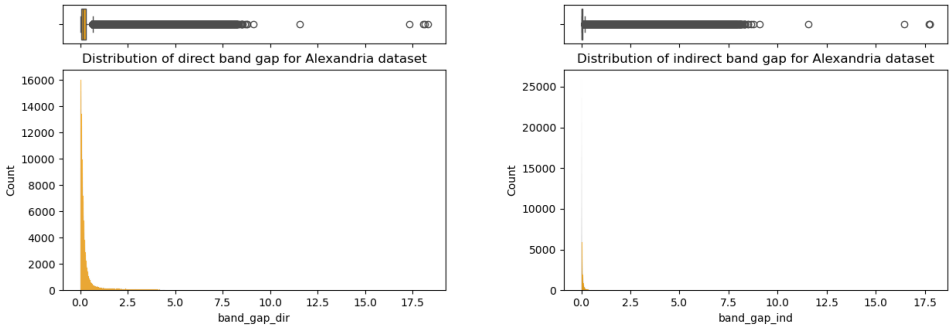


(a) Distribution of Conduction band minimums in the MP dataset



(b) Distribution of Valence band maximums in the MP dataset

Figure 6.3: VBM and CBM



(a) Distribution of direct band gap in the Alexanderia Dataset

(b) Distribution of indirect band gap in the Alexanderia Dataset

Figure 6.4: Band Gap

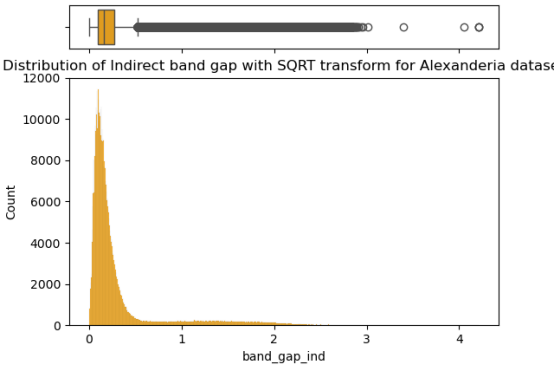


Figure 6.5: Distribution of indirect band gap with square root transform

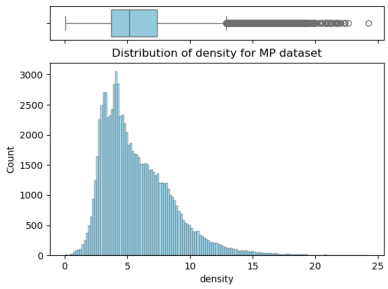
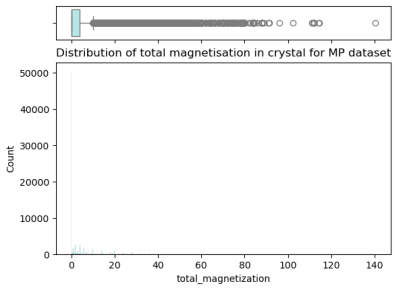
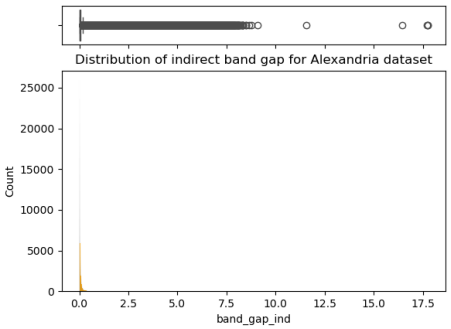


Figure 6.6: Raw distribution of density in Materials Project. It displays a shape common for a Poisson distribution.



(a) Distribution of Total magnetisation



(b) Distribution of indirect band gap in the Alexanderia Dataset

Figure 6.7: Band Gap and total magnetisation

Chapter 7

Appendix B - Results

7.1 Models

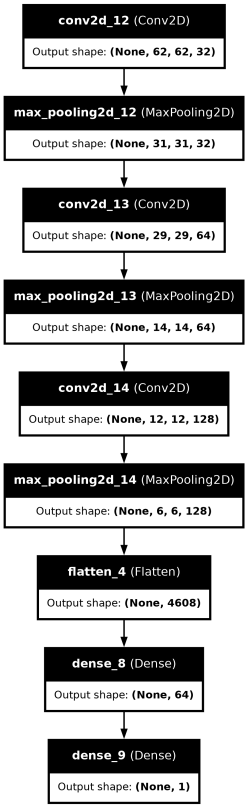


Figure 7.1: Architecture of the baseline model for property prediction

7.2 Generated Data

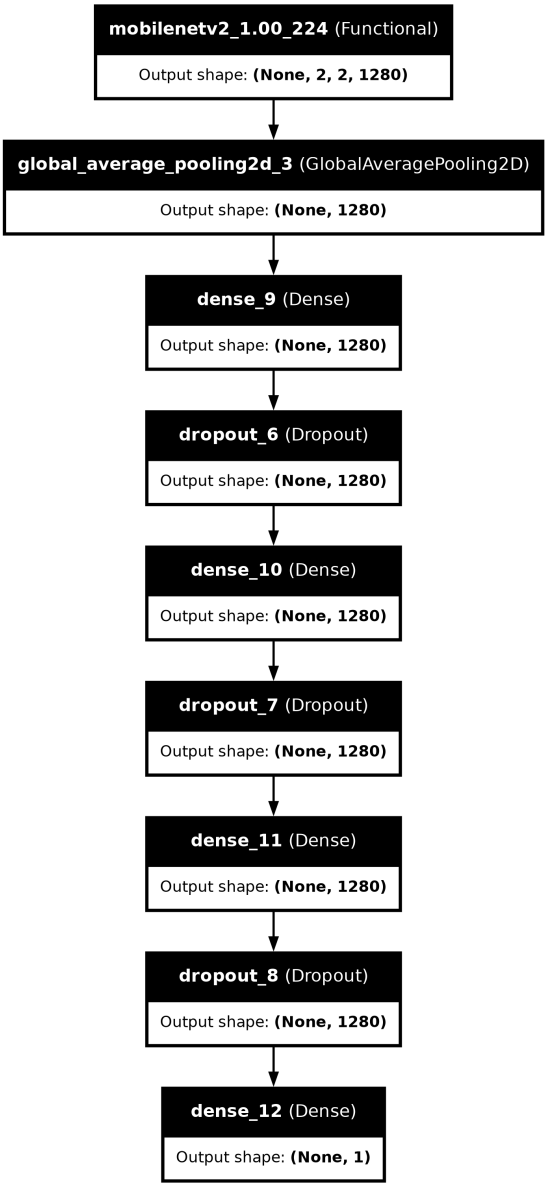


Figure 7.2: Architecture of the pre-trained model for property prediction

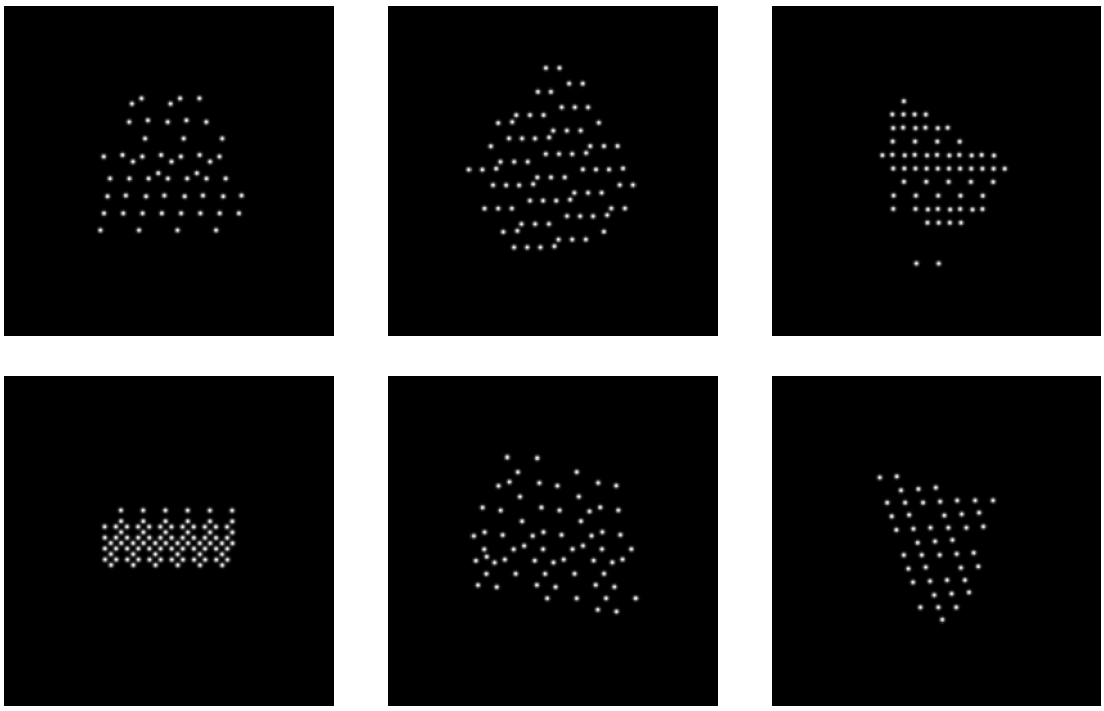


Figure 7.3: An assortment of ECD images from the Imagen model. Note the level of similarity between these and the training images given below.

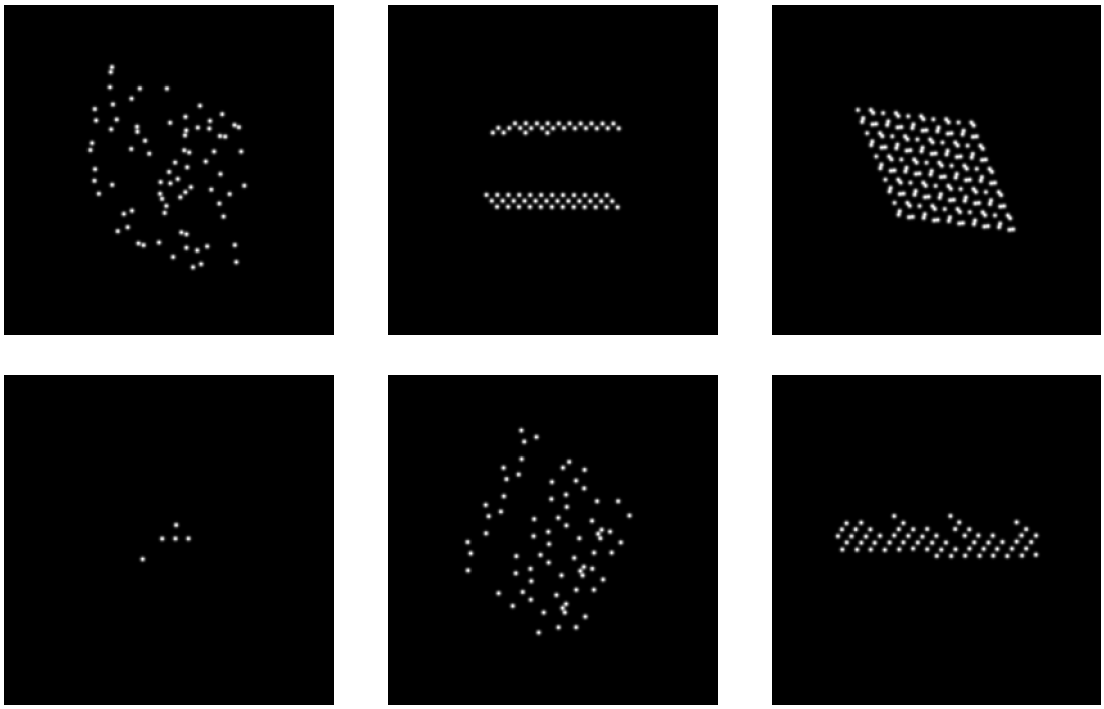


Figure 7.4: An assortment of ECD images from the data augmentation strategy which was used to train an Imagen Model. Compare these with the 6 randomly selected images generated by that model, found below.