Prediction of Battery Materials Properties with Machine Learning

Developing algorithms to discover electrodes for Li-ion and Mg-ion batteries

Sondre Torp

Thesis submitted for the degree of
Master in Materials Science for Energy and Nanotechnology
60 credits

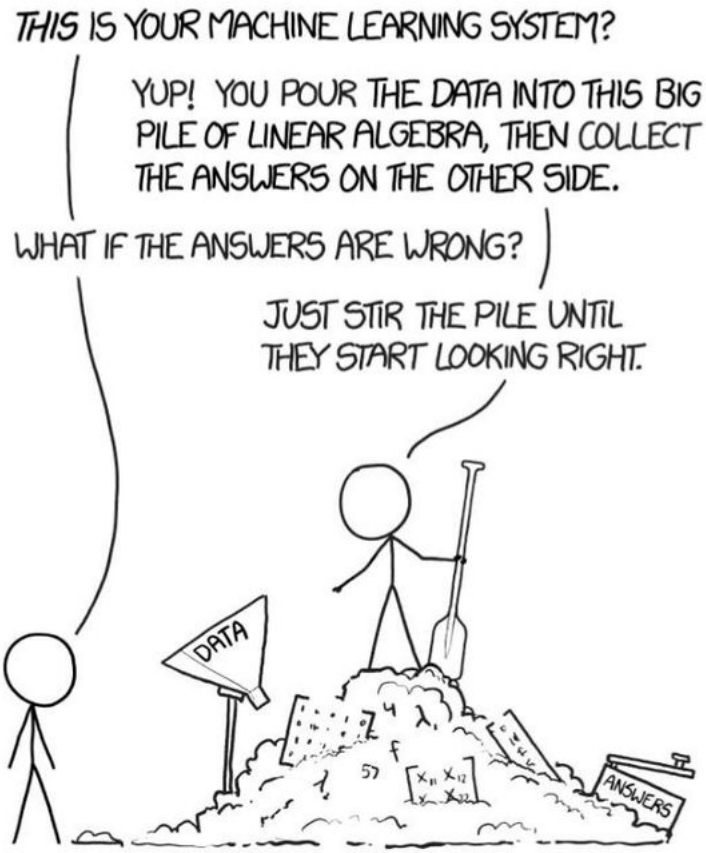
Department of Chemistry
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Spring 2020

Prediction of Battery Materials Properties with Machine Learning

Developing algorithms to discover electrodes for Li-ion and Mg-ion batteries



Prediction of battery materials properties with machine learning - Developing algorithms to discover electrodes for Li-ion and Mg-ion batteries

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

This is the Abstract!

Acknowledgements

This thesis is the result of two years of work under the supervision of prof. Sabrina Sartori at the University of Oslo, in the period August 2018 to March 2020.

First, I would like to thank my supervisors Dr. Sabrina Sartori, Tor Svendsen Bjørheim, Dr. George E. Froudakis and Dr. George Fanourgakis. Sabrina, thank you for all the help throughout my masters thesis, your inexhaustible patience and resourcefulness when we discovered a new problem. Tor, thank you for insight regarding DFT and thoughts on direction. George Froudakis, thank you for welcoming me to Greece with open arms and a open heart. Thanks to you I will always feel at home on the island of Crete. Lastly, thank you George Fanourgakis, you have been one of the greatest contributor for my success, and thank you for sharing your wast ocean of knowledge on chemoinformatics.

The University of Oslo and the University of Crete are filled with people that have helped me. I thank you all. Especially, Professor Morten Hjorth-Jensen for igniting my interest for computational physics during his courses at UiO, and later when discussing my thesis. Also, thank you Manolis Tylianakis for always welcoming me with a smile at UoC and being the first to propose a solution.

I would like to thank all the friends that I have made over the past years while studying at the university. Especially, I would like to thank Blindern Studenterhjem for being a great place for students, and all the good friends I have made there.

Lastly, I would like to thank the people that gave me good laughs and comfort in this period, motivated me to work harder, and helped me, both with spell checking and fixing the errors that was relevant for the work. Thank you Embla Maria O'cadiz Gustad, Mikael Kiste, Mari Røsvik and Erik Lund. Thank you Ida Grøn

Roepstorff and Erik Skaar, your help has been immensely appreciated.

Abbreviations

AP-RDF Atomic property weighted radial distribution function

API Application Programming Interface

AV Average Voltage

CIF Crystallographic Information File

db Database

DFT density functional theory

ED Energy Density

EV electric vehichles

GC Gravimetric Capacity

HT high-throughput

LCO $LiCoO_2$

MAE Mean Absolute Error

MD molecular dynamic

ML machine learning

MOF Metal Organic Framework

MP Materials Project

MSE Mean Square Error

msp material specific properties

PC Principal Component

PCA Principal Component Analysis

RDF Radial Distribution Function

RF Random Forest

RMSE Root Mean Square Error

RQ Research Quesiton

SE Specific Energy

SS Sum of Squares

std Standard deviation

SVR Support Vector Regression

UFF Universal force field

VC Volumetric Capacity

vf Void Fraciton

vnd Volumetric number density

WAPE Weighted Absolute percentage error

Contents

Part I

Introduction

1 Overview

This work focus in Machine Learning (ML) applied in the field of batteries. Specifically, a new method is developed to predict chemical properties like voltage, capacity and energy density of chosen electrode materials.

In the introduction a general background for this work will be given. The chapter includes a motivation, the scope of the thesis and its structure.

1.1 Motivation

The motivation for this work originates from the rapidly increasing demand for improved batteries, both for vehicular and stationary applications, with longer life, lower cost, and adequate energy storage options.

Batteries are vastly complex and much effort have been devoted to their development, in recent times [1] [2] [3], and even 3-D printed batteries are on the rise [4]. Yet, with all the efforts put in to electrochemical cells, there is still a never ending chase for batteries that can push the limits of their properties even further. The demand for better batteries is growing faster than ever. The global electric car fleet, for instance, is exceeding 5.1 million, almost doubling the number of new electric car registrations in the last year. According to the EV30@30 Scenario [5] the aim is to reach a 30% market share for electric vehicles (EV) in all models except two wheelers by 2030. This is because more than one quarter of global greenhouse gas emissions comes from this sector alone. The EV sales per year are then predicted to be more than 43 million sold EVs, and the stock numbering more than 250 million EVs. It is clear that millions of new EVs will push the demands on the battery technology sector [6], with the market requiring high capacity and high energy density batteries.

Some of the most important cell properties are; voltage, energy density, specific energy or capacity, flammability, available cell constructions, operating temperature range, shelf life or self discharge, low cost, and worldwide consumer distribution. Most of these properties are to an extent dictated by battery chemistry.

Due to the complexity of the chemical processes involved, it is of high importance to be able to develop predictive modeling methods to search for better compositions and performance. In this work a method has been developed, to predict; voltage, energy density, specific energy, and the physical stability of materials as electrodes.

Today some of the main methods for theoretical advances in battery science are density functional theory (DFT), molecular dynamic (MD) simulations and machine learning. In the field of computational materials science, a large amount of both theoretical and experimental data has been generated during the last couple of decades. This is, in large part, due to the success of DFT, MD simulations and the increase in computational power. These methods combined with the high-throughput (HT) approach have generated a lot of data and made it, in cooperation with big projects like the Materials Genome Initiative, easily available. DFT is a cornerstone for simulation procedure in materials science [7] [8] [9], while MD simulation is, among other things, known to be well suited to explore solid-state materials at the atomic level [10] [11] [12] [13].

The experimental and traditional approaches to improving battery technology come with a high cost and time-consuming procedures of synthesis. Different applications are usually locked to one type of material because of the investment associated with large-scale production. A change of material is thus rare, and can be considered a revolution. This is why the success of the initial material selected in one sector is crucial for that technology's lasting success. Many new inventions, with close following niches of technologies, demand the development of their own set of materials with properties tailored to that specific technology. Properties from compatibility to toxicity are essential and make the search for materials a multi-dimensional problem [14].

Traditional computational methods, like DFT and MD, come with a high computational cost compared with machine learning methods. ML uses past data to find relations and correlations between the data. Based on them, the models (ML models) created can be used for prediction in new unknown materials. For a ML model to be accurate there are two requirements; a large set of data, often referred to as "big data", and that the model is given the right *descriptors*, where descriptors can be thought of as representations, individual measurable properties or characteristics, of a compound. Big data, in the form of material databases al-

ready exist, like Materials Project (MP) [15], AFLOWLIB consortium [16], OQMD [17], NOMAD [18], and others [7].

The advantages of the ML approach are several - it is computationally efficient - it only takes a few minutes/hours to build a model, and seconds to make predictions. One example of a ML approach towards the search of novel materials is exemplified in Sendek et al., where their ML approach took < 1 second per prediction, while their DFT approach took approximately four weeks per prediction, on solid Li-ion conducting materials [19].

Another advantage of a ML approach is that no mathematical or physical relation, and no laws of nature are needed for good predictions. The ML model will find these relations, but for most of the models these relations are difficult to interpret. In principle, a ML algorithm can be given any information (input). If some of the information is irrelevant, then the ML algorithm will give zero weight to that information. However, if only one part of the relevant information is given to the ML algorithm, the model will not give good predictions.

There are a couple of challenges related to the use of ML models. The need for big data, and a sufficient amount of *examples*, are needed. The accuracy of said data is important, since without accurate data there are no accurate predictions. Descriptors need to be formulated in such a way that the ML model understands the input. Lastly, the choice of ML method is of importance, i.e. Random forest, Support Vector Regression (SVR), Neural Networks, etc.

In real physical or chemical problems, the ML method can provide good estimations, but these are not exact predictions. There are several reason for this, i.e. small training examples, accuracy of the training data, not providing all relevant descriptors, etc. However, a good ML model can provide us with the most promising materials for a given application, or as a minimum, significantly reduce the number of material candidates that experimental studies should focus on.

1.2 Scope of the thesis

This work aims to develop a methodology to predict selected battery properties accurately without the need of large scale simulations, or computer heavy calculations. Using state-of-the-art machine learning, and properties taken from the existing database Materials Project, this thesis propose a set of predictors to discover the properties of new, not yet explored electrodes, or even new properties in already well known electrodes. The properties used as targets for our predictions are; average voltage, gravimetric and volumetric capacity, specific energy, energy density and the stability of the materials. More specifically, this work looks at the theoretical values of the given targets, and not experimental values.

The main objective of this thesis is to acquire knowledge about which features (i.e. descriptors) should be included for electrode predictions of the given targets by investigating a range of possible inputs, their configurations and their effect on prediction accuracy. It examines both features found in the Materials Project database, and uncommonly used features in literature. To the best of our knowledge, a general examination of possible features for electrodes is lacking in research. Knowledge of which features improve prediction accuracy is useful in several ways. Even though material data generation is done faster than ever before, data acquisition and formatting can be time consuming, especially if an abundance of data is used. It can therefore be of great value to know what type of data should be prioritized, therefore, giving focus to the areas we know will improve our predictions. This saves time from both data mining and testing. Knowledge about which properties required can also expose both, the effectiveness of less obvious features (that can be overlooked by other developers of ML-models), and the need for not yet introduced characteristics. Reducing the number of features required will also reduce the complexity of the model, which saves energy and time. This work tries out different sets of features which should be considered when doing predictions on electrode materials.

1.2.1 Research Question

Specifically, this work seeks to answer the following questions:

RQ1: Is there potential for the of use machine learning to improve the search for good battery materials?

RQ2: What predictors are suited for such a task?

RQ3: Do features overlap? What features should be removed from the feature space to achieve the most efficient training?

RQ4: How does the size of the database affect the results?

RQ5: Which ML method would be the most optimal for such a search?

1.2.2 Approach

RQ1 regards the need to establish a database that is comprehensive enough to be of value for the selected machine learning method, and it must have the relevant information needed to get good predictions.

RQ2 concerns the choice of features examined in this work. They are inspired by a survey done on a similar project in the field of Metal Organic Framework (MOF) performed by collaborators from the university of Crete [20] [21], and another research project done on MOFs by Fernandez *et al.* [22]. The choice of predictors were also, to a degree, dictated by the lack of more data, the difficulty of finding said data, and by the data we had.

Regarding what descriptors applied: First, physical descriptors such as geometrical properties (volume, number of sites, type of atoms, *etc.*) of the unit cell, were tested. It was greatly efficient in similar studies on MOFs, and it is straightforward. Thereafter, *void volume* seemed like a good candidate due to the nature of intercalation type batteries, and void volume is both a geometrical feature, and it is computationally cheap to obtain. Geometrical dependent properties gave predictions that were too inaccurate. A more chemical approach, without doing large DFT type calculations, were needed. An atomic property weighted radial distribution function approach were tested to include listed values like electronegativity, van der waals volume and magnetization. Before lastly, using the already known results (or targets) to make predictions on the remaining targets.

RQ3 examines the evaluation of the features space and how to limit its size. To achieve the most effective training principal component analysis was applied, thus systematically removing the redundancy within features. This is a statistical procedure that uses an orthogonal transformation on the data to make a set of linearly uncorrelated variables and rank these after variance.

RQ4 concerns the size of the database. All of the features were tested alone, and up against each other, both for a smaller database of Mg-ion intercalation batteries, and a bigger database of Li-ion intercalation batteries. The results of these two different datasets might not be purely comparable, but due to the algorithm mainly using fundamental properties the results will be compared and trends examined.

RQ5 is approached as follows: Which machine learning method is more likely to yield correlation in the dataset?

1.3 Structure of the thesis

In this section a short presentation of how this thesis is structured is given.

Part II: Foundation

Chapter 2: Batteries

In this chapter a short history of batteries and their evolution is given, before going deeper into the state-of-the-art of Li ion batteries. An introduction to the theory of battery cells, their operation principles and design, is given. Lastly, the battery properties of particular interest and their features related to this work, are presented.

Chapter 3: Machine Learning

This chapter introduces the field of machine learning along with some key concepts. Challenges concerning the application of machine learning are also discussed. The subgroup of machine learning algorithms, ensemble methods is introduced in the context of decision trees with a special emphasis on Random forest.

The evaluation methods used in this work (e.g. root mean square error deviation, K-fold cross validation, etc.), as well as principle component analysis are presented. Finally, an introduction to state-of-the-art computational material design with an emphasis on electrodes and battery related works, is given.

Part III: Experimental method

Chapter 4: Method

This works relies on data from the Materials Project database. These datasets are introduced in this chapter, as well as the features used for the analysis, and a brief touch on the preprocessing of the data. Lastly the experimental environment is explained with the prediction pipeline. you gotta do this

Part IV: Results and discussion

Chapter 5: Results

This chapter presents and discusses results for RQ1, RQ2, RQ3 and RQ4 with experiments performed for both databases and predictors on all targets. The first section of the chapter investigates the predictors one by one, and compare their results on the different databases. All calculations are available on github.

Part V: Summary

Chapter 6: Conclusion and future work

In this chapter the most important findings from this research are revisited and put into perspective. In addition, suggestions for future work are laid out.

Part II

Foundations

2 Batteries

This section presents a brief summary of some of the major steps in the history and evolution of batteries, with a description of lithium and magnesium-ion batteries and their role in todays market. The basic principles of batteries will be explained, with a special emphasis on electrodes. Lastly, some of the chemical properties related to this work will be introduced.

2.1 History and evolution of batteries

One of the main issues regarding the development of sustainable and clean-energy technologies are the lack of efficient energy systems [14]. A tremendous amount of resources are used on an international level to produce batteries with higher capacity, voltage and energy density. The evolution of batteries started in Italy with Alessandro Volta (1745 - 1827). He built the first known battery in the year 1800 [23]. His invention consisted of the voltaic pile, with zinc and copper plates stacked on top of each other and sheets of brine-soaked cardboard between each plate. The revolutionary property of the voltaic pile was that it could produce a stable current for longer periods of time, not just short sparks of electricity. This invention was the foundation of todays modern battery (figure 1).

Almost 40 years later the British inventor John Frederic Daniell Volta continued this line of work, with the discovery of the Daniell cell [25] in 1836. The Daniell cell, as illustraited in figure 2), is constructed with two half cells, one with a zinc electrode in a zinc sulfate dissolution, and a copper electrode in a copper sulfate solution. These half cells are connected by a salt bridge. The cell could give a voltage of $1.1\,\mathrm{V}$ through the reaction shown below 2.1.

$$\mathrm{Zn_{(s)}} + \mathrm{Cu^{2+}_{(aq)}} \rightarrow \mathrm{Zn^{2+}_{(aq)}} + \mathrm{Cu_{(s)}} \tag{2.1}$$

In 1859 the French physicist Gaston Planté built the first led-acid battery.

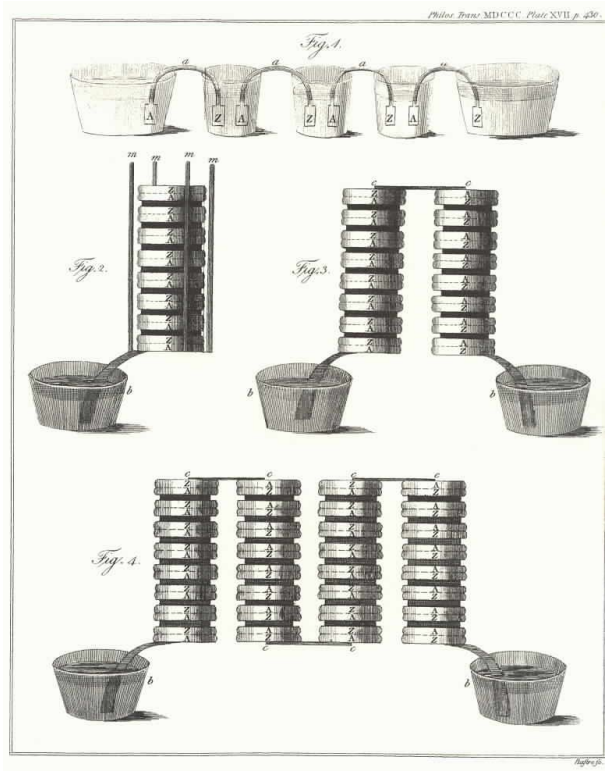


Figure 1 – A voltaic pile, the first battery [24]

The battery could be charged by applying an external opposite potential, and it was the first secondary battery every made. Planté rolled two led plates into a spiral, separated by rubber strips, so that the plates would not touch. The lead-acid battery was special due to the electrolyte being an active part of the chemical reaction. The electrodes were led anode, and led(IV)oxide cathode, immersed in sulfuric acid. The overall reaction is shown beneath 2.2. Both the anode and the cathode are made into led(II)sulfate during discharg. The charge is depleted in the electrolyte when the battery is completely discharged (the sulfuric acid has a lower density). Charging changes the electrolyte back into concentrated sulfuric acid.

$$\mathrm{PbO_{2(s)} + Pb_{(s)} 2\,H_2SO_{4(s)} \rightarrow\ 2\,PbSO_{4(s)} + 2\,H_2O_{(l)}} \tag{2.2}$$

The open circuit voltage (V_{OC}) (i.e. the voltage between the terminals with

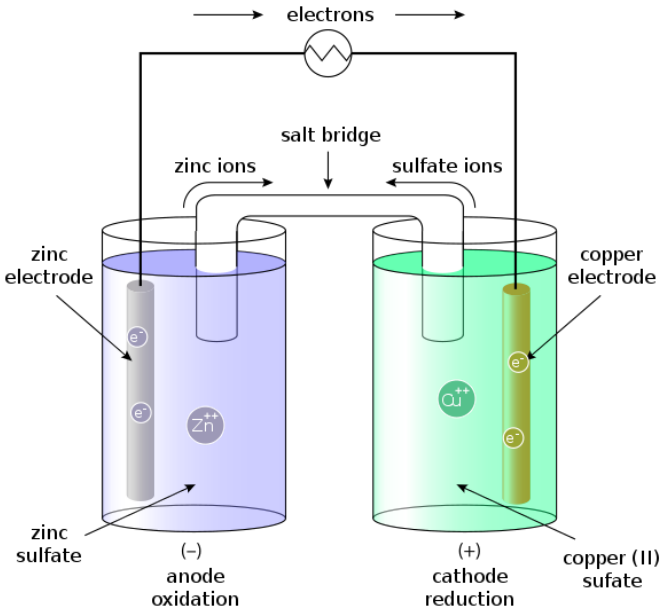


Figure 2 – A draft of a Daniell cell. The anode is a piece of zinc and the cathode a piece of copper. The salt bridge transports ions between the solutions and the electrons moves through an external circuit [26].

no load applied) for a led-acid battery is approximately $2\,\mathrm{V}$. It is custom to attach these batteries in series to attain a higher voltage, typical $6\,\mathrm{V}$ or $12\,\mathrm{V}$. These devices have a shelf- and cycle-life of more than 10 years or $1,000 - 2,000$ cycles. They are still being used in modern cars. Led-acid batteries have a relative low specific energy, which means that the current is low compared to their weight another drawback is their high environmental impact. Therefore, one of the many goals of battery producers is to replace led-acid batteries with higher performing alternatives.

Nickel-cadmium (NiCd) batteries were first described by the Swede Waldemar Jungner in 1899 [27]. These batteries rose in popularity due to their high energy density, low weight, long shelf-life, and their relative fast recharge. Typically, they yield a nominal cell voltage (i.e. the reported or referenced voltage) of $1.4\,\mathrm{V}$. The cathode is made of Nickel oxide hydroxide, the anode of metallic cadmium, while the alkaline electrolyte is a basic solution of potassium hydroxide. The specific energy of a typical Nickel-cadmium is $40 - 60\,\mathrm{Wh/kg}$. Equation 2.3 shows the overall reaction of such a battery.

$$2\,\mathrm{NiOOH} + \mathrm{Cd} + 2\,\mathrm{H_2O} \rightarrow 2\,\mathrm{Ni(OH)_2} + \mathrm{Cd(OH)_2} \tag{2.3}$$

Nickel-metal hydride (NiMH) batteries were first commercialized in the 1980's and had various similarities with the NiCd batteries. The main difference is the anode which is replaced by an alloy of metal hydrides (MH). NiMH batteries have the same electrolyte as NiCd batteries, a solution of potassium hydroxide. The nominal cell voltage of such a battery is typically around $1.2\,\mathrm{V}$ and the specific energy is $60 - 120\ \mathrm{Wh/kg}$. Equation 2.4 shows the overall reaction of a NiMH battery.

$$\mathrm{Ni(OH)_2} + \mathrm{M} \rightarrow \mathrm{NiO(OH)} + \mathrm{MH} \tag{2.4}$$

A primary cell is a non-rechargeable battery. These batteries are usually used in remote controls, flashlights, and other small household appliances. Alkaline manganese batteries, or just alkaline batteries, are one of the most common primary cells in modern society, with anodes of zinc, cathodes of manganese oxide and an electrolyte of potassium hydroxide. A typical alkaline battery delivers a nominal cell voltage of $1.5\,\mathrm{V}$. The overall reaction for an alkaline battery is shown below (2.5).

$$\mathrm{Zn} + 2\,\mathrm{MnO} + \mathrm{H_2O} \rightarrow \mathrm{ZnO} + 2\,\mathrm{MnO(OH)} \tag{2.5}$$

2.2 Lithium based batteries

The intercalation electrodes for lithium and other alkaline metals were discovered in 1975 by Michael Stanley Whittingham [28]. This led to the first lithium batteries with titanium disulfide $(\mathrm{TiS_2})$ as the cathode and metallic lithium as the anode. $\mathrm{TiS_2}$ - structure consists of layers where lithium-ions are inserted or

extracted without significant changes in the structure, which makes the reaction reversible. Figure 3 shows the layered structure of TiS_2. During discharge of the battery, lithium-ions leave the anode of metallic lithium and moves through the electrolyte and into the empty octahedral position in the TiS_2-structure, while titanium(IV) reduces to titanium(III). Applying an over-potential to the material charges it and lithium-ions move out of the TiS_2-structure with titanium oxidizing back to titanium(IV). This discovery was the start of a major research investment in cathode materials of sulfite and other chalcogens in the 70-80's.

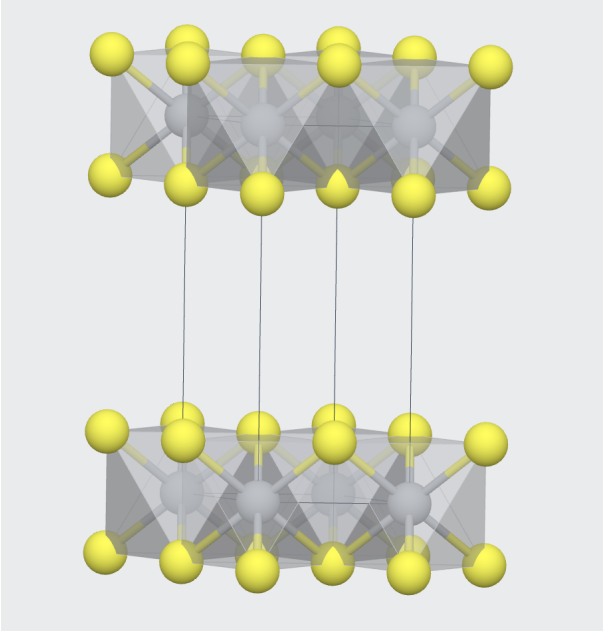


Figure 3 – The two-dimensional structure of TiS_2. From a slight angle along the b-axis. The titanium in grey, sulfur, in yellow. Lithium-ions would intercalate into the space between the TiS_2 layers [29].

The layered structure of these type of electrodes allow their reversible behavior. In 1980 John B. Goodenough introduced $LiCoO_2$ (LCO) as the cathode material for lithium batteries. This earned him together with M. Stanley Whittingham and Akira Yoshino the Nobel Prize in Chemistry in 2019 [30]. Goodenough and colleagues obtained a current density of up to $4\,mAcm^{-2}$ [31] [32]. Even though the properties where exceptionally good at the time, the batteries were still not commercialized due to metallic lithium being too unstable, ergo an unsafe anode material. This was due to dendrites growing out of the anode that short circuited the battery.

In 1991 Sony introduced lithium batteries, with LCO as the cathode, on the commercial market. LCO compounds provide good electrical performance, are relatively safe, easy to prepare, and are not especially sensitive to process variation and moisture. The metallic lithium anode was substituted for graphite which reduced the growth of dendrites at the anode. The electrolyte was an organic solvent with a lithium salt.

A lithium-ion battery refers to a battery where lithium intercalates in both electrode materials, both the cathode and the anode. Lithium batteries, instead have an anode of metallic lithium. This nomenclature is transferable to other type of batteries like magnesium-ion/magnesium batteries.

Figure 4 shows a typical lithium-ion battery with $LiCoO_2$ as the cathode and graphite as the anode. During discharge the lithium-ions move from the anode, through the electrolyte and separator to the cathode. The electrons move from the anode to the cathode through a separate external circuit, where the electrical energy can be extracted. When charged an over-potential is applied and the reaction is reversed. The overall reaction is shown in equation: (2.6)

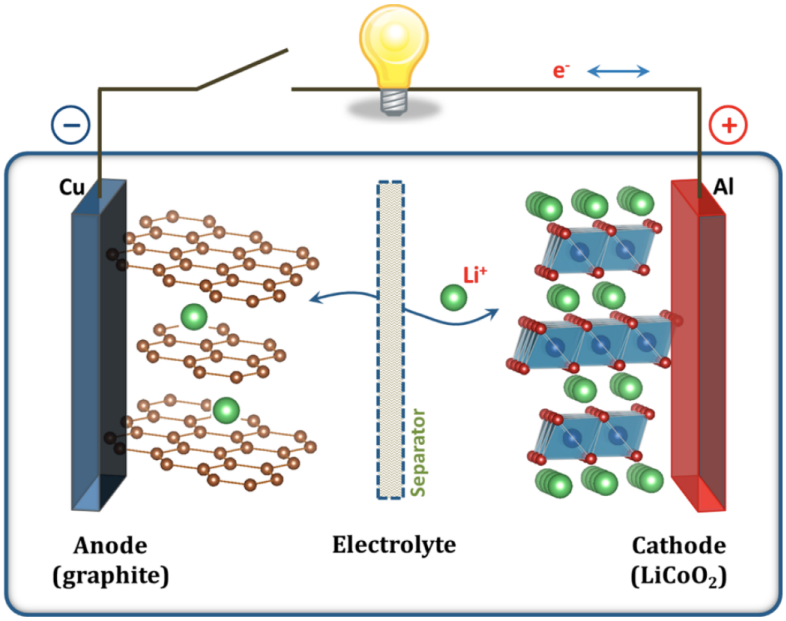
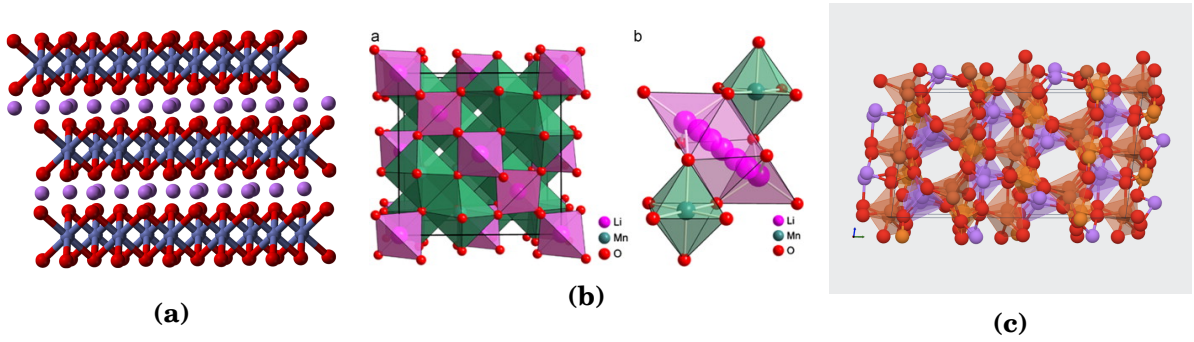


Figure 4 – Schematic illustration of the first Li-ion battery $LiCoO_2/Li^+$ electrolyte/-graphite [33].

$$LiCoO_2 + C_6 \rightarrow Li_{1-x}CoO_2 + Li_xC_6 \tag{2.6}$$

The cathode materials used in lithium-ion batteries have evolved since the 1990s. Typical cathode materials, as of today, are $LiMn_2O_4$ (spinel) and $LiFePO_4$. $LiMn_2O_4$ is a good ionic conductor due to the structure having channels in all three dimensions where lithium can be transported 5b. $LiFePO_4$ has the lower ionic conductivity of the two, due to only having channels in one dimension, as shown in figure 5c. Even with a lower ionic conductivity, it is still a popular material due to its long cycle life.



(a) **(b)** **(c)**

Figure 5 – Crystal structures of the layers in a) $LiCoO_2$, [34] b) the 3-dimensional channels in $LiMn_2O_4$ [35], c) and the 2-dimensional channels in $LiFePO_4$ [36] are illustrated.

The most used anode materials for lithium-ion batteries are graphite and other forms of carbon based materials. Graphite has a high energy density, making the cathode material the limiting factor for energy density of Lithium-ion batteries. The improvement of the cathode material is therefore high in priority among many research groups. Another recent anode materials is $Li_{4/3}Ti_{5/4}O_4$ spinel, which has a lower specific capacity than graphite, but has a longer cycle life and good thermal stability characteristics. Nanostructured $Sn-Co-C$ alloys commercialized in 2005 by Sony and Si-based negative electrodes seem promising for Li-ion cells with higher specific energy and energy density.

The main reasons for the use of Li-ion batteries can be summarized as follows: They have a long shelf and cycle life, low self discharge rate, high energy efficiency,

high energy density, high rate and power discharge capabilities, no memory effect and many possible chemistries offer design flexibility. While some common drawbacks are: moderate initial cost, degeneration when discharged below $2\,\mathrm{V}$, degradation at high temperature (above $65^{\circ}\mathrm{C}$ they can permanently lose capacity), their need for protective circuitry, capacity loss and potential for thermal runaway when overcharged and when crushed. Some also become unsafe if rapidly charged at sub zero temperatures.

For more than 40 years, the search for batteries with efficient energy storage, high capacity, long cycling and shelf life has been necessary to satisfy our demands for cheap, transportable power. Lithium batteries using lithium metal anode have attracted attention due to their promises of high energy storage capacity. However the batteries are prone to dendrites when plated, which results in short circuit and fire hazards [37][38], Many possible solutions are being proposed [39] [40] [41] [42]. In recent years a desire to move towards an ultimate energy density technology has forced researchers to evaluate technologies beyond Li-ion batteries, where other metals such as magnesium and aluminum are pointed out [43] [44]. Aluminum and magnesium are considered because of their abundance. In the case of aluminum it has a high theoretical voltage, a high specific energy, and it is the most abundant metal in the world. It is hindered by a oxide layer on its surface [45], but solutions to this problem is being offered for large-scale applications [46]. Aluminum-based batteries are outside the scope of this work and will not be discussed.

There is also an ongoing search for candidates for solid-state electrolytes, due to energy density and safety being the main factors that govern the development of the rechargeable battery technology [47]. Solid-state electrolytes would enable stable and reliable operation of all-solid-state Li-, Na-, and Mg-based batteries. Special focus is given to lightweight complex metal hydrides, due to their high ionic conductivity, and in some cases electrochemical properties that enable battery reversibility.

2.3 Magnesium based batteries

Magnesium batteries have been used as a primary battery, but historically, there has been little interest due to hydrogen gas generation during discharge, and relatively poor storage-ability of partly discharged cells. When fully charged the storage-ability, even under high temperature, is good [43], which has made the battery relevant for military application.

Recently magnesium batteries have attracted increased attention due to Mg higher volumetric capacity than lithium (i.e. $3832\mathrm{mAhcm}^{-3}$ vs $2061\mathrm{mAhcm}^{-3}$). Being the fifth most abundant element [48], makes magnesium, with its low atomic weight, low cost, and electrochemically active nature, a good candidate for battery applications. It can serve as a possible negative electrode with its electrochemical potential of $-2.37\,\mathrm{V}$, and it is environmentally friendly.

While not competitive with Li metal on both specific capacity ($2205\ \mathrm{mAhg}^{-1}$ vs $3862\ \mathrm{mAhg}^{-1}$) and redox potential ($700\ \mathrm{mV}$ lower), dendrite formation is absent, which alleviates the safety concerns [49]. Still, there are several roadblocks ahead when looking at the possible electrolytes. One is the unique electrochemistry which prohibits its reversible deposition in aprotic solvents contained in commercial ionic salts such as magnesium bisimide or magnesium perchlorate. Magnesiums low reduction potential gives it a tendency to form surface films that hinder ionic conductivity, opposite to Li compounds who also creates surface films, but these being ionic conductors and behaving like solid electrolyte interfaces. This not being the case for Mg compounds which creates blocking surface layer that inhibits deposition and conduction of magnesium ions [50] [51].

There is an ongoing search for high performance cathode magnesium materials for the realization of a practical, rechargeable Mg battery. The Mg^{2+} shows promises of a instant multiplication of the electrical energy that can be released for the same volume, but the strong interaction between the Mg^{2+} ions and the host create problems [46]. This have lead to a search for electrodes and electrolytes that will allow the double charged magnesium ions to move through the host more easily. It is almost two decades since the first secondary magnesium battery was made, but these batteries are still at the research stage [52].

2.4 Cell operation principles and design

Batteries are electrochemical devices. They store chemical energy that can be converted into electrical energy. This is done by an oxidation-reduction (redox) reaction where one of the species in the reaction gains or loses an electron by changing the oxidation number. One battery consists of one or more *cells*. A cell is fundamentally made of three parts; the anode, the cathode, and the electrolyte. The anode is a negative electrode, which refers to the direction of current through the electrode. It is commonly a metal that would oxidize if given the opportunity. For a conventional current flow the electron moves from the anode to the cathode. The anode is often low voltage. The cathode is a positive electrode. The cathode is a metal that is normally combined with oxygen and is where the reduction occurs. A common example of a oxide is iron oxide. The cathode is normally high voltage. The electrolyte is the material that, when introduced to the anode and the cathode, provides an electrically conducting medium for transfer of charge. Electrolytes are typically liquid, to impart the ionic conductivity. It can be a solid, but this is, at least for now, less common. The cell will produce electricity when the circuit is complete. The electrolyte can, in some designs, act as both electrolyte and anode or cathode.

If the anode is made from pure metal and has an external cathode of ambient air it is referred to as a metal-air electrochemical cell. These batteries have a much higher theoretical energy density. However there are technical issues confronting their development. [53]

The difference between high- and low voltage is referred to as the cell voltage, which is the driving force for the discharge of the battery. For secondary batteries, it is possible to recharge batteries by reversing this process by applying an external electrical power source, it creates an over-potential, i.e. a higher voltage than the one produced by the cell, with the same polarity.

Changes in the design of the cell dictates the cells performance. If the compositions of the electrodes are altered, the cell will yield a different amount of electricity. Adjustments in the cell can affect the amount of electricity, the rate of production, the voltage, and the cell's ability to function in different temperatures. There is almost an endless amount of possibilities, even though the most common

cell has been 1.5 volt alkaline batteries. Other types of batteries include Lithium batteries, Magnesium batteries, Zinc batteries, Mercury batteries and others.

2.5 General introduction to battery properties

In a cell there are essentially two areas, or sites, in the device where the redox reactions occur. In general these half-cell reactions can be expressed as one reduction and one oxidation reaction:

$$\mathrm{aA} + ne \leftrightharpoons \mathrm{cC}$$

Where a is the number of molecules of substance A taken up by n electrons to form c molecules of C, and the oxidation reaction defined in the same way:

$$\mathrm{bB} \leftrightharpoons \mathrm{dD} + ne$$

with the overall reaction, as exemplified by the Daniell cell 2.1 being:

$$\mathrm{aA} + \mathrm{bB} \leftrightharpoons \mathrm{cC} + \mathrm{dD} \tag{2.7}$$

Whenever there is a reaction there is a decrease in the free energy of the system. This is free energy is called standard Gibbs energy and it is defined as:

$$\Delta G^0 = -nFE^0$$

Where n is the number of electrons in the reaction, and F is the Faraday constant ($F = 96485\ \mathrm{Cmol}^{-1}$). Gibbs free energy of the reaction is the driving force of the battery and enables it to deliver energy to an external circuit. E^0 is the standard potential of the cell. It is determined by the type of active material in the cell, i.e. the difference in electrode potential between the cathode and anode.

E decides how easy it is to remove one electron from a atom. It can be calculated from the free energy or from the standard electrode potential 2.8.

$$\text{oxidation potential} + \text{reduction potential} = \text{standard potential} \tag{2.8}$$

e.g. from our database:

$$\mathrm{Li^+(aq)} + e^- \rightarrow \mathrm{Li(s)} \qquad -3.04\,\mathrm{V} \tag{2.9}$$

$$\mathrm{Bi^{3+}} + 3\,\mathrm{e^-} \leftrightharpoons \mathrm{Bi(s)} \qquad 0.317\,\mathrm{V} \tag{2.10}$$

$$E^0 = 3.357\,\mathrm{V} \tag{2.11}$$

Direct measurements of the absolute electrode potential is difficult to achieve, so a reference point is defined. The standard potential of H_2/H^+ is set to zero and all other standard potentials are compared to this potential. If to metals are interconnected in a electro chemical cell, the metal with the larger standard reduction potential will gain electrons. A rule of thumb is, from low to high, alkali metals, alkaline earth metals, aluminum, base metals (e.g. Fe, Ni), hydrogen, and transition metals.

In situations where the system is not in the standard state, the *voltage* E of a cell is given by the Nernst equation.

$$E = E^0 - \frac{RT}{nF}\ \ln \frac{a_C^c a_D^d}{a_A^a a_B^b} \tag{2.12}$$

where a_i is the activity of the species. R is the gas constant, and T is the absolute temperature.

The **voltage** can be defined as the difference between two electrical potentials. In most batteries the electrical potential difference occurs due to the redox reaction in the electrodes that creates a potential gap between the electrode and the electrolyte. When an outer circuit is connected this gap is lowered, but due to

the reaction rates going up, the potential gap is maintained.

Capacity is a measurement of how much charge a battery can hold. It is most common to evaluate the capacity of an electrode or a battery in terms of capacity per weight $\mathrm{mAh/g}$, i.e. the gravimetric capacity (GC). It can also be denoted as capacity per volume $\mathrm{mAh/m^{-3}}$, i.e. the volumetric capacity (VC). The capacity of a battery is often compared to the theoretical capacity which is determined by the amount of active material in the cell. It can be found by Faraday's law 2.13.

$$Q = \frac{mFz}{M} \tag{2.13}$$

Where F is Faraday's constant ($F = 96485\ \mathrm{Cmol^{-1}}$), z is the valence number of ions of the substance, m is the mass and M is the molar mass of the substance in grams per mol.

Capacity can also be defined as:

$$C = \int I(t) \cdot dt \tag{2.14}$$

Where i is the number of electrons or cations exchanged between the negative and positive electrodes, i.e. how much charge a battery can store. $I(t)$ is the current, the number of electrons flowing over the external circuit per time interval dt, which is integrated over the discharge period. Theoretically, capacity is 1 gram equivalent weights of the active material (in grams) divided by the number of electrons in the reaction.

If the calculations are based on only the active materials participating in the electrochemical reaction the theoretical capacity of a $\mathrm{Zn/Cl_2}$ cell is $2.54\mathrm{g/Ah}$ or $0.394\mathrm{Ah/g}$.

$$\mathrm{Ze} + \mathrm{Cl_2} \rightarrow \mathrm{ZnCl_2} \tag{2.15}$$

$$1.22\ \mathrm{g/Ah} + 1.32\ \mathrm{g/Ah} = 2.54\ \mathrm{g/Ah} \tag{2.16}$$

Notably, when calculating the theoretical capacity of a battery it is higher than the actual capacity. This is due to the mass of the electrolyte, separator, and other battery components that is missing from the equation.

The active materials of the electrodes allow the reversible uptake and release of ions. This may happen by movement of the ions in a couple of different ways. They can move into their chemical structure through intercalation, or they can move out of their chemical structure, through extraction or deintercalation. Lastly, this can also be done by conversion of the electrode material into other more ion rich/poor chemical forms or mixtures.

The total Li or Mg content in the electrodes will either be varied by changing the composition of one phase or the ratio between coexisting phases. In this work we will only look at intercalation type batteries, as will be discussed.

The voltage of a battery determines the work a battery can do and depends on the types of active materials used. The cell voltage is also limited by concentration and temperatures, as expressed by the Nernst equation. A higher voltage is desirable, because of the increase in work that can be done by the battery.

The calculation of the voltage of a lithium ion battery is more complex than calculating the voltage of a common electrochemical cell with two electrodes in a wet solution. The voltage of an electrochemical cell is calculated from the difference in chemical potential for the lithium on the anode and the cathode [54], as shown in eq2.17.

$$V_{OC} = \frac{\mu_A - \mu_C}{F} \tag{2.17}$$

Where F is Faraday's constant and μ_A is the chemical potential of the lithium anode, and μ_C is the chemical potential of the lithium cathode. The cell potential is thus decided by both the difference in electronic potential and the lithium ions movement. The energy from the electronic potential is calculated from the redox potential of the lithium cathode and anode, while the energy from the ion movement is dictated by the crystal structure and the coordinates of where the lithium ions were intercalated or deintercalated.

Energy density (ED) is related to the capacity of a battery. The energy density of a material is the energy of a system per volume ($\mathrm{mWh/g^{-1}}$). Another closely related term is specific energy (SE), which is the energy per unit mass ($\mathrm{J/kg}$). The formula for energy density is given in equation 2.18

$$P = Q \cdot U \tag{2.18}$$

Where P is the efficiency or energy density of the material. Q is the capacity of the material, and U is its potential. We can calculate the energy density for a battery with $\mathrm{LiCoO_2}$ as the cathode and a graphite anode (while ignoring the rest of the battery) as shown in equation 2.19, assuming a average voltage of $3.6\,\mathrm{V}$.

$$P = 100\ \mathrm{mAhg^{-1}} \cdot\ 3.6\,\mathrm{V} = 360\ \mathrm{mWhg^{-1}} \tag{2.19}$$

This is the theoretical energy density of a battery with LCO as the cathode and graphite as the anode. The specific energy density, i.e. where the battery is included in the calculations in its entirety, of the same composition results in a energy density of $190\ \mathrm{mWhg^{-1}}$ [55].

Capacity and energy densities of battery materials can be compared relative to mass, volume and cost. The more electrode material that a battery contains, the greater is its capacity and energy. The higher the cell voltage the greater its power and energy.

Some relations important for this work are the relation between energy, energy density, capacity, power and current. These relate to each other as shown in equation 2.20, where E is the energy (Wh), V is the voltage (V), C is the capacity (Ah), U is the energy density ($\mathrm{J/m^3}$), P is the power (W), I is the current (A) and t is the time (h)

$$V \cdot C = E \tag{2.20}$$

$$\frac{E}{Volume} = U \tag{2.21}$$

$$V \cdot I = P \tag{2.22}$$

$$W \cdot t = E \tag{2.23}$$

Energy is the cells ability to do work, which is a property of high interest for practical applications.

The battery can deliver power which is defined as:

$$P(t) = V(t)I(t) \tag{2.24}$$

Where $I(t)$ is defined as earlier, and drawn at a cell voltage $V(t)$. The amount of work that can be done by the battery or the energy contained in the battery, is then defined as the power delivered over the discharge period.

$$W = \int P(t) \cdot dt = \int V(t)I(t) \ \cdot dt \tag{2.25}$$

This is particularly interesting for applications that require a lot of work in a short time period.

2.6 Cell definitions used in this work

In this thesis, especially under the section on general properties of battery, terms related to features used from Materials Project database are used. These terms will be defined or clarified here.

The features discussed here are based on optimal design and discharge conditions. These values are helpful to set a number on the "goodness" of a battery. The actual performance may vary under normal conditions of use.

Energy is the computed energy, it is the total energy or sum of the electronic energy and nuclear repulsion energy.

Energy per atom is the computed energy normalized per atom in the unit cell.

The *formation energy per atom* is calculated from the formation energy from the elements normalized per atom in the unit cell.

Volume is the volume of the unit cell.

Band gap is the distance between the valence band and the conduction band, it represents the minimum energy that is required to excite an electron up to a state in the conduction band. In general, band gaps computed with common exchange-correlation functionals such as the LDA [56] and GGA are severely underestimated [57]. Typically the disagreement is reported in luterature to be $\tilde{5}0\%$. Internal testing by the Material Project supports these statements; reporting band gaps underestimated by $\tilde{4}0\%$.

Density, here defined as the calculated bulk crystalline density. Typically underestimated due to the calculated cell volume being overestimated on average by $3\%(\pm 6\%)$ [58].

Magnetic moment (μ_{B}) is calculated for the unit cell within the provided magnetic ordering.

Number of sites is the total number of atoms in the unit cell.

Elasticity is the predictor associated with the elastic properties of a solid, i.e.

the elastic constant. It provides a complete description of the response of the material to external stresses in the elastic limit [59].

Polarizability is a tabulated atomic properties. It is the ability to form instantaneous dipoles, and is defined as:

$$\alpha = \frac{P}{E} \tag{2.26}$$

Where α is the polarizability in isotropic media, p is the induced dipole moment of an atom to the electric field E that, is the field that produces the dipole momentum.

Van der waals volume (V_W) of a molecule are the space occupied by the individual atom, which is impenetrable to other molecules at ordinary temperatures. For a single atom, it is the volume of a sphere with a radius equal to its van der Waals radius (r_W):

$$V_W = \frac{4}{3}\pi r_W^3 \tag{2.27}$$

Physical stability is the energy above hull in eV. It is the energy that is demanded for decomposition of the material into the set of most stable materials at that chemical composition. Positive values indicate that the material is not stable. While a zero energy above hull indicates that this is the most stable material at its composition. Stability is tested against all potential chemical combinations that result in the material's composition. For example a $\mathrm{Mg_3Sb_2}$ structure would be tested against other $\mathrm{Mg_3Sb_2}$ structures, against Mg and Sb mixtures, and against MgSb and $\mathrm{Sb_2}$ mixtures.

In a battery, the reactant is supplied from the electrolyte phase to the catalytic electrode surface. Electrodes are often composites made of active reactants, binders and fillers. To minimize the energy loss of both activation and concentra-

tion polarizations at the electrode surface and to increase the electrode efficiency, it is often preferred to have a large electrode surface area. This can be done by having a porous electrode design. A porous design can provide an interfacial area per unit volume that is considerable higher than that of a planar electrode.

A *porous electrode* is an electrode that consists of porous matrix of solids and void space. The electrolyte penetrates the void space of a porous matrix. In such an active porous mass, the mass transfer condition in conjunction with the electrochemical reaction occurring at the interface is very complicated. In a given time during cell operation, the rate of reaction within the pores may vary significantly depending on the location. The distribution of current density within the porous electrode depends on the physical structure (pore size), the conductivity of the solid matrix and the electrolyte, and the electrochemical kinetic parameters of the electrochemical processes.

3 Machine Learning

In this chapter we summarize some concepts of machine learning and related concepts. The first section introduces the basic ideas behind machine learning and one of the best known examples will be presented. Secondly the concepts of supervised and unsupervised learning will be presented with a clarification regarding the difference between regression and classification problems, so that we can discuss where in the field of machine learning this work resides in. Basics of methods utilized in this work will be introduced, emphasizing Random forest. Subsequently a short description of the validation methods used is given. These are; K-fold cross validation and how it is used in optimizing our random forest method, root mean square error (RMSE) and R-squared(R^2).

We conclude this section with a brief explanation on the role of data, how features can affect the effectiveness of a model, of the concepts of over- and underfitting, and how these are related to the bias-variance-trade-off.

Sondre: Did you forget something? Come back to this when done with the section.

3.1 The basics of Machine Learning

Machine learning comes from the field of pattern recognition and learning theory, and is defined as the field of study that gives computers the ability to learn without being explicitly programmed. Or more precise: "... A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E..."([60]). At its core the ability to learn by detecting patterns in usually huge amounts of data that, more often then not, is impossible to perceive for a human.

3.1.1 The basics

As an introduction on how machine learning was applied to learn and recognize patterns in our work, it will be useful to start with a simple example applied to the recognition of the handwritten number "5".

How two people writes a single digit may vary to an extensive degree. It might seem to be a easy problem, but if the recognition is to be done manually million of times, it is no longer a trivial task for any human being. Therefore a model which can recognize these digits would be useful. A model that takes a picture of a digit as input and outputs that digit in a way that is recognizable for a machine, i.e. a digital format.

Machine learning only works when you have data, preferably a large amount of data. For instance data from the MNIST test dataset [61]. This database contain $60,000$ images of handwritten numbers that is commonly used for both various training, and testing in the field of machine learning. The images all are 18x18 pixels. The data is divided into two sets, one training set: X_{Train} and one test set: X_{test}. Some numbers from the MNIST database are shown in figure 6.

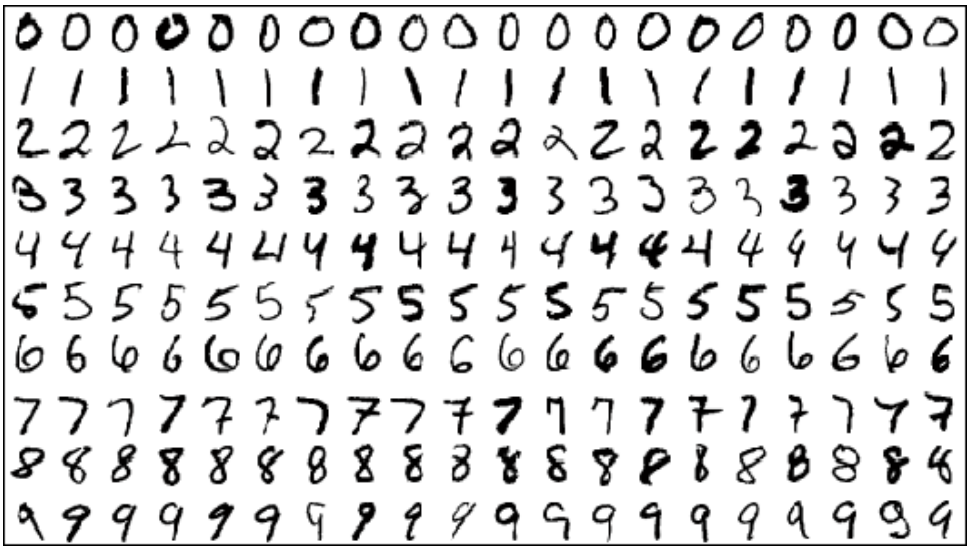


Figure 6 – Number from the MNIST database [62]

How do one represent an image as something that makes logical sense to a computer? Most learning algorithms take numbers as input. To a computer one image a grid of numbers that represent how dark a pixel is. So each picture con-

tains a gray-scale value that ranges from 0 to 255. Where each sample can be viewed as a vector consisting of 324 *features*. Every sample has a corresponding label value, or *target*, which is the digital equivalent to the handwritten sample. We let the corresponding targets be denoted: y_{train} and y_{test}, for training and testing data. Next we designate our *learner* denoted by function h. h is then given our training set S, where $S = (X_{train1}, y_{train2}), ..., (X_{trainN}, y_{trainN})$ and returns a prediction rule: $h : X \ \rightarrow y$. This rule is also called a predictor, in general, a classifier, or a regressor, depending on the problem in question.

The *training phase* is a process where the learning algorithm gets tweaked to best capture the correlating structure of the data set, so that it can better predict new data. As mentioned in the last paragraph the output from the *training phase* is called a *predictor*. The next step is to introduce the *predictor* for new, unseen data, so that it can be classified. Then we compare the y_{test} to our predicted value y_{pred} given by h to see if our model generalizes well to unseen data in X_{test}.

3.1.2 Supervised and Unsupervised Learning

One of the most basic separations in machine learning is the partition between supervised learning and unsupervised learning [63].

In the case of supervised learning, the answer to a problem is known and given to the computer. The computer then deduce its own logic to figure out how to get to that result, thus the name complete-date problem is commonly used. This is the most common type of learning. With unsupervised learning the machine is tasked with finding patterns and relationships in data sets without any prior knowledge of the system, incomplete-data problems. Some authors operate with a third and a forth category, namely reinforcement learning, where the machine learns by trial-and-error, and evolutionary learning, where they account for the biological evolution and that it can be seen as a learning process [64].

In this thesis, only supervised learning is considered. Algorithms and challenges specifically related to unsupervised learning, reinforcement learning, and evolutionary learning, is therefore not further examined.

3.1.3 Regression and Classification Problems

A response variable can either be qualitative or quantitative in nature. For the qualitative response variable, let's assume a set of data points $\vec{x}$ and a goal of finding the value of the output y when $x = 0.5$. The value x is not in the data points given so a way to *predict* the value, is needed. Given in the example above, we assume that there exists a function h that the value comes from. When that function is found one can find any given y for any given x. This is what is known as a regression problem - The response variable takes form of a continuous numerical value. The regression problem is a problem of function approximation or interpolation. It may occur a scenario where there are multiple functions, lets say h and g, that fits the given data perfectly. If this is the case one value in-between the data points is selected, and both the functions, h and g, tries to predict its values and the results are compared to see which is better. This does not seems as very intelligent behavior, but the problems of interpolation can be very difficult in higher dimensional space. This can also be observed in classification, the other aspect of what our algorithms can do.

If the response variable is quantitative the problem is referred to as a classification problem. Such a problem consists of taking several input vectors and deciding which of N classes they belong to. This decision or prediction comes from training on examples of each class. To be clear, classification problems are of a discrete nature - The input only belongs to one class, like the example given at the start of this section 3.1.1.

In this work we want to predict characteristics of batteries, these properties have continuous values, meaning that the task at hand is a regression problem.

3.1.4 Data collection, Preparation, Features and Feature Selection

Normally the data collection is a large part of the work and not easily available, or, at the very least, needs to be assembled and prepared. If the problem is completely new it might be natural to engulf this step with the next one. (Which is, more or less, what this work tries to do.) With a small dataset with many differ-

ent features one can experiment and try to figure out what features are the most useful before picking those and collecting a full dataset based on them and then perform a complete analysis.

A common problem in similar studies is that there are too many types of data that can be relevant, but that data is hard to find or represent in a way that makes sense for the machine. This can be because it requires too many measurements, or, something that is prevalent in this work, that they are in a variety of places and formats. For instance; if the measurements are already taken, but at vastly different temperatures they might be hard to compare or merge. It is important to have a *clean* dataset, this means that the dataset does not have missing data, significant errors, and so on. On top of all of this, supervised learning requires a target y, which demands time and involvement of experts.

The specific input to a model is normally referred to as a feature, that is, numerical representation of raw data. The amount of features are of importance for the machine learning algorithm to successfully make a good prediction. If there are to few relevant features one can not make an accurate prediction due to the lack of necessary data. And if there are to many features, or many of the features are irrelevant to the task the model will be more expensive.

The amount of information needed is extensive, and should be of high quality. A bigger dataset demands a higher cost, and predicting the amount of data required is a futile endeavor. Luckily Machine Learning is still less computationally costly than modeling full systems at a micro or nanoscale, which makes it interesting in the field of material science.

3.2 Bias-variance tradeoff

As the algorithm learns we need to make sure that it generalizes well to data not in its training set. Obviously the algorithm can not generalize beyond the limits of the training data. Therefore it is important to minimize the two sources of errors known as *bias* and *variance*. This is know as the *bias-variance trade off*. It is the property of trying to minimize the two errors simultaneously, and should not be confused with the *irreducible error* of a model which is a result of the noise of the

data. These three together are the terms used to analyze an algorithms expected *generalization error*, which is a measurement of how accurately an algorithm is able to predict outcome vales for unseen data.

Our machine is bias if it generalizes too much. The error is due to low variability in our training data, or that it did not adapt to the training data appropriately. The machine misses the relevant relations in the data set between the features and the output. This effect leads to that which is commonly referred to as under-fitting, see left on figure 7.

Variance is the error that stems from high variability, and the degrees of variability in most machine learning algorithms is large [64]. In simple terms, there is a low degree of generalization. It might be a perfect fit, but as soon as new data are introduced the predictions plummet. This is commonly referred to as over-fitting, see right on figure 7.

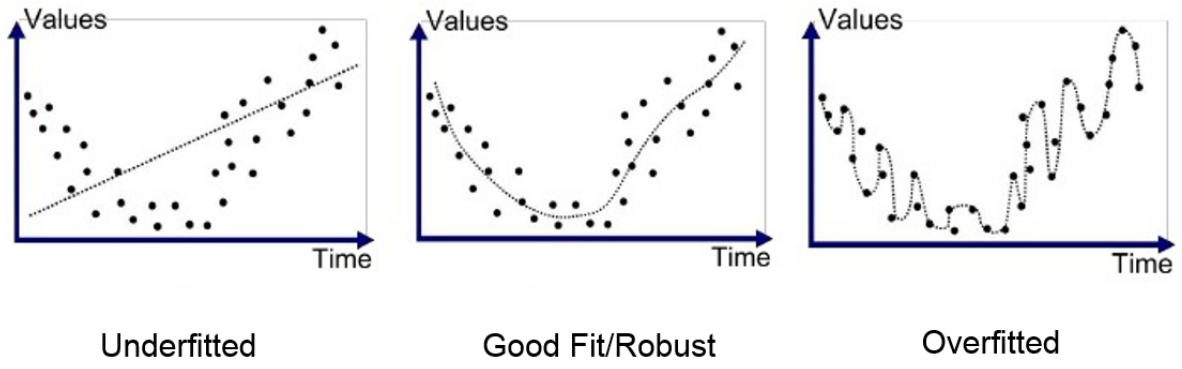


Figure 7 – Simplified illustration showing the concepts of bias-variance problem. Left to right; high bias, low bias and low variance, high variance 7

.

A good way to understand the idea of the bias-variance tradeoff is, a more complex model with an increased number of features is not necessarily better at predicting what you want to predict.

3.3 Random Forest

3.3.1 Ensemble learning

There are many different machine learning algorithms, in this work we have focused on the *ensemble method*; *Random forest* [65]. The idea of ensemble learning is that two heads are better than one, so why not have many learners that all get slightly different results on the same data, and then combine them, as shown in figure 8.

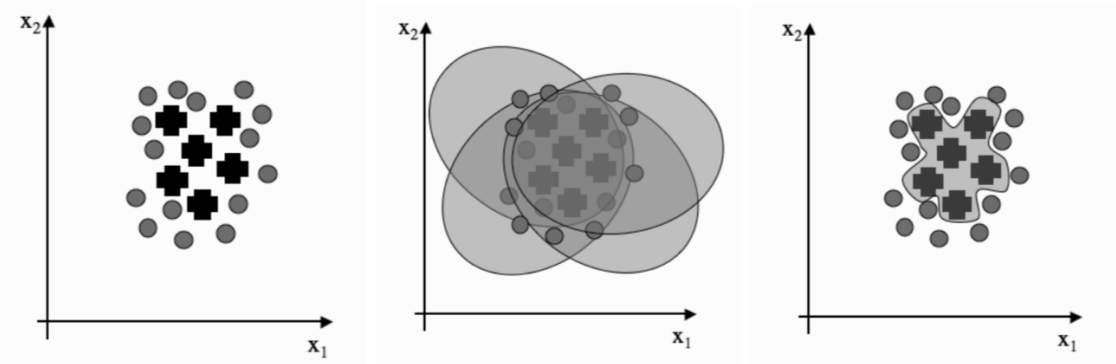


Figure 8 – Combining different classifiers trained on the same data, which in combination can make a much better decision boundary on the target data. Adopted from [64]

Ensemble methods are particularly useful in machine learning when there is little data, as well as when there are to much data, this is heavily due to cross-validation, which will be explained later 3.4.8).

3.3.2 Decision tree

A decision tree is a low cost binary flowchart-like structure. It is one of the most common data structures in the field of computational science, both because of the low cost to make the tree, but also because the cost of using the tree is even lower; $\mathcal{O}(\log N)$, where N is the number of data points [64].

Decision trees are structured much like a regular tree 9, at the top there is a

base, or a *root*, down the branches there are chance nodes, and at the end of the branches there are *leaves*, or end nodes. Every internal node is structured like an conditional statement on a feature.

Let us say that you want to play tennis. You look out the window, and there are three possible weather states (root node); rain, overcast or sun. If it is overcast you will play either way, but if it is windy, you need to evaluate if that wind is strong or weak (chance node). If it is little wind, you will play. Else, it is strong and you will not play (end nodes).

The chance nodes are the results from these tests, and the leaves are the class labels. The full route from root to leaf is the classification rule, or *branch*. An advantage of random forest being based in decision trees is that the algorithm is much more like a "white box" compared to Neural networks "black box" approach, because we can retrace the decisions of each tree. This is especially helpful in the research done in this work where we want to figure out the roll of every feature, and how they affect the result.

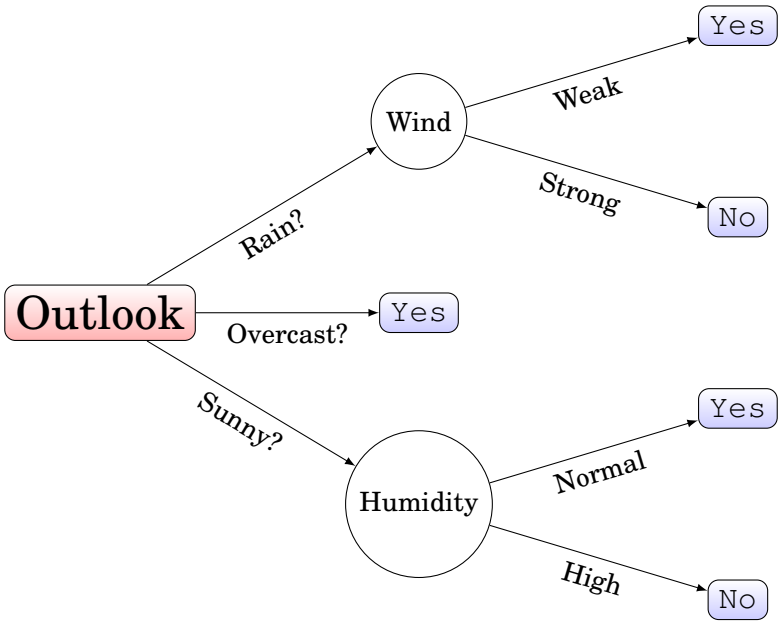


Figure 9 – A simple example of a decision tree for playing tennis. Root in red, leaf node in blue.

3.3.3 Random forest

Random forest (RF) is a ensemble learning method, the idea is that one decision tree is good and many trees, or a forest, is better. The most interesting part of random forest is the randomness that is introduces. Several classifiers are achieved by using the simple combination method *bagging*. Bagging stands for *bootstrap* aggregating. Bootstrapping is the process of taking a sample from the original dataset at random, and replacing parts of it with other original data, so that it is not equal to the original data. There will then be several samples where some of the data is equal, while others are completely different. For bootstrapping in random forest, one sample is taken from the dataset for each tree.

A new parameter is then introduced, at each node a random subset of features are given to the tree, it can only make decisions based on that specific subset, and not the original tree. This increases the randomness in the creation of each tree, and it speeds up the learning process. The reason to add randomness to the algorithm is to reduce variance without effecting bias. It also removes the need for decision tree *pruning*, i.e. reducing the complexity of decision tree by removing the parts of the tree that does not help the classifier and it reduces overfitting. The process of creating trees is repeated until the error stops decreasing.

When the forest is done, we use a majority vote system, which is a comparison of the mean response for regression. For a step by step algorithm, see the README.txt file on github. The reason for not using cross-validation in the learning algorithm, which is common in other machine learning methods 3.4.8, is that our bootstrap method only uses about 65% of the data, leaving 35% on average which can give a estimated test error.

The main reason we decided to opt in for random forest is due to an article by [66] and the findings from both our collaborators [67] and Shandiz and colleagues [68], that clearly state that random forest is the chosen machine learning algorithm when you want to test for correlations. Another reason and a main advantages of RF is that it is fast and does not require any particular optimization of its hyper-parameters (e.g. number of decision trees for RF). On the other hand, methods like support vector regression (SVR) require an extensive search for the optimum hyper-parameters before providing reasonable results.

3.4 Evaluation method

3.4.1 Mean and Variance

One of the most recognized properties of a distribution is its *mean*, or expected value. It is denoted by μ, and defined as: $E[X] = \sum_{x \in \chi} xp(x)$, for discrete variables.

The variance is a measure of the "spread" of a distribution, denoted as σ^2. It is defined as:

$$\text{var}[X^2] = E[(X - \mu)^2] \tag{3.1}$$

from which:

$$E[X] = \mu^2 + \sigma^2 \tag{3.2}$$

can be derived. For a ML model, a prediction is done with an accuracy x_acc on training data and its prediction accuracy on test data is y then:

$$\text{var} = x_{acc} - y \tag{3.3}$$

3.4.2 Standard deviation

Standard deviation (std) is a tool to quantify the measure of spread. It is very similar to variance by yielding the measure of deviation whereas variance provides the squared value. Standard deviation is defined as:

$$std[X] = \sqrt{var[X]} \tag{3.4}$$

or:

$$std = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N-1}} \tag{3.5}$$

3.4.3 Mean square error

The Mean Square Error (MSE) can give a measure of the quality of our estimator [69]. It is defined as

$$\text{MSE}(\epsilon) = \frac{1}{n}\sum_{n}^{n-1}\epsilon^2 = \frac{1}{n_{\text{samples}}}\sum_{n}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2 \tag{3.6}$$

Where $\hat{y}_i$ is the predicted value of the i-th sample, and y_i is the corresponding true value. As such it can be thought of as the average of the square of our residuals. Therefore the MSE can never have negative values, and smaller values mean that we have a better prediction, where at zero there is a perfect fit.

3.4.4 Root mean square deviation

The Root mean square deviation, or root mean square Error (RMSE), is defined as the square root of the MSE:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{n}^{n-1}(y_i - \hat{y}_i)^2}{n}}$$

And is thus the distance, on average, of a data point from the fitted line, measured along a vertical line. The RSME is directly interpretable in terms of measurement units, and is therefore a better measure of goodness of fit than a correlation coefficient.

RMSE and Stdev can seem very similar, but they are not the same. Stdev measures the spread of data around the mean, while RMSE measure distance between some values and predictions for those values. If the mean error approaches 0 and n approaches infinity Stdev and RMSE converge.

3.4.5 Mean absolute error

Mean absolute error (MAE) is another statistical tool that is used to measure the difference between two continuous variables, in our case; the predicted values and the observed values. It corresponds to the expected values of the absolute error loss. The MAE is defined as:

$$\text{MAE} = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \tag{3.7}$$

where y_i and $\hat{y}_i$ are defined as above. In geometrical terms, it is the average absolute vertical/horizontal distance between each point in a scatter plot and the $Y = X$ line.

3.4.6 Weighted absolute percentage error

Weighted absolute percentage error (WAPE) is the mean absolute error divided by the mean ($\bar{y_i}$) multiplied by a hundred. This yields the mean error in percentage.

$$\text{WAPE} = \frac{\text{MAE}}{\bar{y}_i} \times 100 \tag{3.8}$$

3.4.7 R^2 score - The Coefficient of Determination

In regression validation the R^2 is the standard when it comes to measuring goodness of fit [70]. In straight terms it is the proportion of the variance in the dependent variable that is predictable from the independent variable.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{3.9}$$

Where y_i are the indexed response variables (data to be fitted) and f_i the predictor variables from the model with $\epsilon_i = y_i - f_i$. The average of the response variables is denoted $\bar{y}_i$. The second term can also be considered as the ratio of MSE to the variance (the $1/n$ factors null each other out in a fraction), or the total sum of squares (SS_{tot}).

If the residual sum of squares (SS_{res}) is low the fit is good. However, this should be compared to the spread of the response variables. After all, if the response variables are all nicely distributed close to the mean, then getting a good SS_{res} is not suspicious. We therefore do a normalization in the fraction, taking the scale of data into consideration. In the simplest polynomial fit, using a zero order polynomial (a constant), our model would just be a constant function of the mean. The sums being equal, returning unity on the fraction and the total R^2 score would be zero. In the other extreme, if the model fits perfectly, than SS_{res} would be zero and the R^2 score would be one. In this sense we have a span of possible R^2 scores between zero and one, from the baseline of the simplest model at zero, and a perfect fit at one. In contradiction to most scores the value can be negative, because the model can get arbitrarily worse, thus giving negative values. The R^2 score is useful as a measure of how good our model is at predicting future samples.

3.4.8 K-fold cross validation

K-folding is a cross validation technique that allows us to generalize the trends in our data set to an independent data set. In this way we can circumvent typical problems like over-fitting and selection bias [70]. The approach for the technique is simple and represented in figure 10. Instead of doing a regression on the entire data set, it is first segmented into k number of subsets, or splits, of equal size (making sure to pick out the variables randomly before distributing them to the subsets).

Now one subset can be chosen to be the 'test' or 'validation' set while the rest of the subsets ($k-1$ of the folds) are the training sets. The desirable regression is then applied on the training set, arriving at some data fitting that is the prediction. From here it is a straight forward process to analyze how well our predicted variables compare to the validation variables, e.g. through the R^2 score function. However, even though the subsets are picked randomly, the validation subset used could potentially not be a representative selection of the entire set. Therefore the process is repeated k times, each time using a new subset as the validation subset. Finally, one can simply calculate the average of the scores to get the predictive power of our model. As an added benefit, since the calculations are done anyways, the average of the predictions can be used as the final fit.

Cross validation techniques are very useful when the gathering of new data is difficult or impossible, as we are using the extra computational power at our disposal to squeeze the most amount of relevant information out of our data.

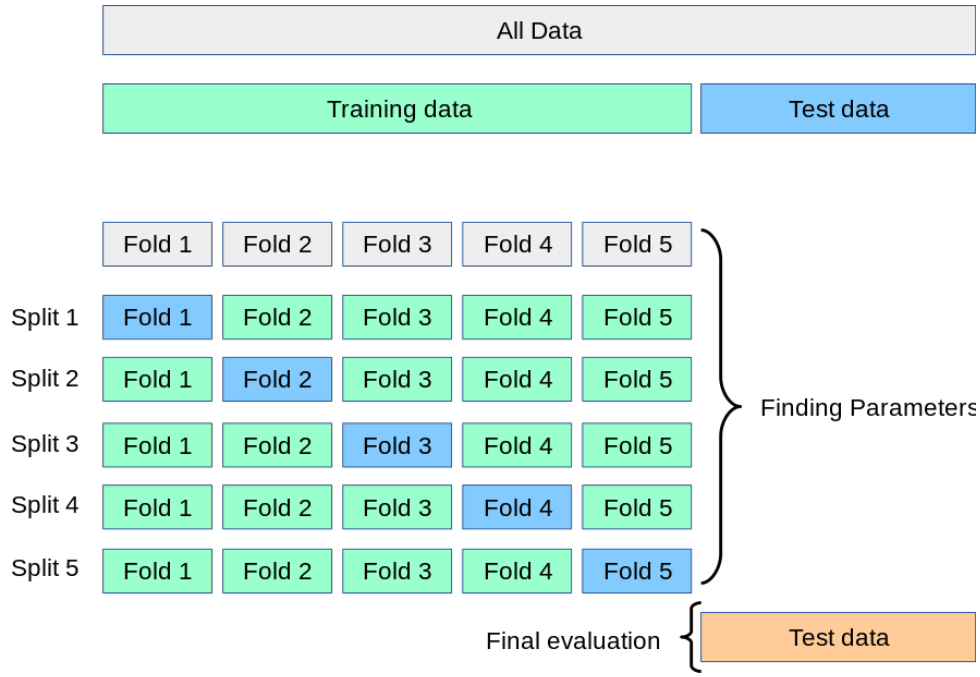


Figure 10 – A representation of how cross-validation is done. First the data is split into K number of sets, then one of these sets are left out as test data. The model trains on the training data before being tested on the test data. This process is repeated K times, and the mean is taken [71].

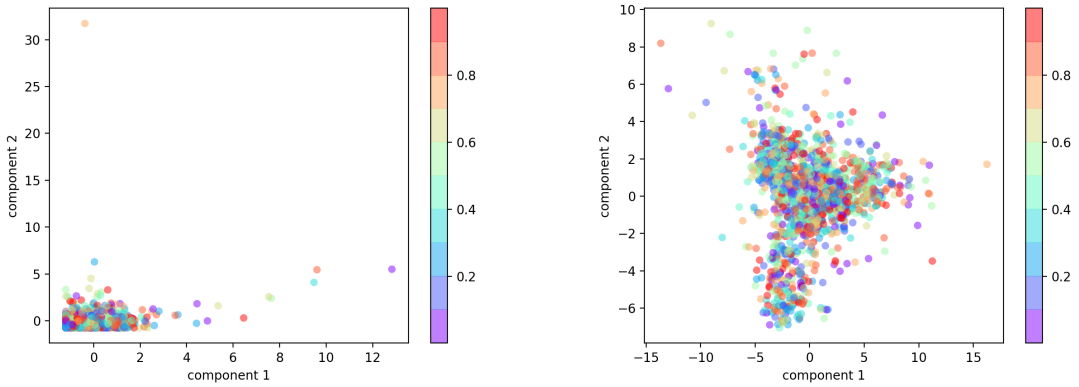
3.5 Principal Component Analysis

Principal Component Analysis (PCA) [70] is a procedure that uses orthogonal linear transformation to reduce the amount of feature subspaces. It goes under different names in different fields, but the most recognizable might be Single Value Decomposition. It consists in converting a set of possible correlated variables into a set of uncorrelated variables, called principal components (PCs).

The PCs are arrayed so that the first PC has the largest variance, meaning that it accounts for the largest possible amount of variability in the dataset. The second PC does the same, it accounts for as much variability as possible with the constraint of being orthogonal to all the former components. These orthogonal vectors are linear combinations of an uncorrelated orthogonal basis set. Graphically, the shortest vectors affect the predictions the least. Since PCA is sensitive to the relative scaling of the original variables, in *sklearn.decomposition.PCA*, (the library we use), the input data are centered but not scaled, before performing the

PCA.

Figure 11 shows, the training data from the Mg-db plotted in two scatter plots. To the left is our original data uniformly scaled, scaled so that the variance is equal to one, the mean of the distribution is zero, and about 68% of the values lies between -1 and 1. To the right, a plot showing the affine transformation of these data (PCA), which have been translated, rotated and uniformly scaled. As can be seen in figure 16a, this affine transformation leads to clearly distinguishable classes.



(a) **(b)**

Figure 11 – Two scatter plots; On the left, some of our data from the Mg-ion database before PCA. On the right, our data after PCA, showing that there are distinguishable classes.

3.6 Earlier work

The field of material science is blooming due to computing power being cheaper and more available than ever before. When the Human genome project started in the 1990 the main problem was lack of computational power, but because of Moore's law, the sequencing of all 3 billion letters of DNA is not even a challenge in todays computational standard. In the field of material science, researchers have known how to simulate materials since the early 1900, due to the discovery of quantum mechanics, the challenges are related to the of computational power.

An advanced material have an order of 10^{23} electrons that demands compu-

tational methods to simplify the problem [72]. The gap between the computation needed and what state-of-the-art systems are providing is increasing [43].

In the field of computational material design a subfield called 'high-throughput' (HT) computational material science is on the rise [73] [4]. This area is based on computational quantum - mechanical - thermodynamic approaches, and a multitude of techniques both in database construction and in the field of intelligent data mining. The idea is simple; first construct a large enough database of accurate thermodynamic and electronic properties of existing and hypothesized materials. Secondly, use different algorithms and statistical models to intelligently analyze the data and find materials with desired properties. This method should continuously be validated by comparing the calculated values with real (already known) materials, and later also on new hypothetical materials, to create a feedback loop to further improve the algorithm [14]. To simplify, a computational HT method consists of three tightly connected steps: Virtual material growth, rational material storage, and material characterizations and selection. This work is based on the second and third step, rational material storage and material characterization and selection, on intercalation type batteries.

Several studies on batteries have applied machine learning and different degrees of HT, in particular to estimate their state of charge accurately and to improve their energy system management [74] [75] [76] [77]. In the direction of predicting battery properties only a handful of studies were found. Shandiz *et al.* [68] used classification methods to determine the crystal system of silicate cathodes. They found that the random forest classifier gave the highest accuracy of prediction, and that there is a strong correlation between the three major crystal system (monoclinic, orthorhombic and triclinic), and other features of cathodes. Their data was taken from the online database Materials project [15] [78] [79]. Sendek and *et al.* [80], constructed a classification model using logistic regression to find possible solid state electrolytes for lithium-ion batteries. They concluded that simple atomistic descriptors alone, were not enough to obtain useful predictions. Instead, using a multi-descriptor model can yield good predictions. Similar to this work Joshi *et al.* [81] employed ML techniques to predict electrode voltage for metal-ion batteries from the Materials Project database. Much like this thesis, their emphasis were on finding proper features vectors that could accurately represent the compounds.

Other areas where ML have shown promise are in the field of Nanoporous Materials [20]. Three a set of new descriptors for predictions on methane adsorption was proposed. Fanourgakis *et al.* combined structural features, such as the helium void fraction, surface area, and pore volume, with other descriptors designed, such as the probe atoms of various sizes on MOFS. The work lead to a more general application of ML on nanoporous materials [21]. It was found that introducing "atom types" as descriptors in the ML algorithm to account for chemical character of both the MOFs and the Covalent Organic Frameworks (COFs) improved the ML predictions significantly.

Part III

Experimental method

4 Method

In this section we will introduce the overall approach to the research. We will first describe the data set and experimental environment, followed by the description of the methods used to represent physical and chemical properties of the electrode materials.

The two most crucial challenges in ML are; creating the dataset, which needs to be large enough and as accurate as possible to improve the predictions. Secondly, finding the right descriptors. Only fulfilling both these conditions, is it possible to create a reliable ML model.

All codes are written in Python 3.7 [82] or Fortran98 [83] and can be found here: github. The main frameworks and libraries we used for our data mining and data analysis are Python with NumPy [84], JSON [85], Pandas [86], and Scikit-learn [87]. Fortran90 was used to run the void fraction and AP-RDF calculations.

4.1 Data set and Experimental Environment

The features used in this work are *volumetric number density* (vnd), *Void fraction* (vf) and *atomic property weighted radial distribution function*(AP-RDF). In the following we mention the basic principles of our algorithm.

Regarding the choice of which database (db) to be considered, as mentioned in the section 1.1 but opted for *Materials Project* [15] [88] (MP),since many useful data were already collected for the *battery explorer*(Battery Explorer) [78] [79]. MP also has a functioning application programming interface (API) [89], further facilitating our work. The db has a sizable amount of information on electrodes. There are $16,128$ conversion electrodes and $4,401$ intercalation electrodes. The dataset contains DFT calculations for Li, Mg, Ca, Al, Zn, and Y intercalation electrodes. We decided to focus the project to the intercalation electrodes of Mg-ion and Li-ion battery type. This left us with two db's; First, a Mg-ion battery db with 355 batteries, accounting for around 10% of MPs intercalation db. Secondly, a Li-ion db with $2,073$ batteries, which is nearly 52% of the entire db. These numbers are

smaller than what is possible to find in MP due to our removal of inconsistencies and repetitions.

The MP db contains the reduced cell formulas with Crystallographic Information File (CIF) files for all voltage pairs, that is; the CIF files for both the charged and discharged material. The different characteristics, or voltage pair properties, are also present. The characteristics that we used as targets in this work, are; Average Voltage (AV), Gravimetric and Volumetric Capacity (GC, VC), Specific Energy (SE), Energy Density ED, and a measurement of the stability. Other properties that where in the MP db and to some extent tested as predictors for each material, both charged, and discharged are: sum of the electronic energy and nuclear repulsionenergy, energy per atom, volume of the unit cell, band gap, density, total magnetization, number of sites, and elasticity 2.5.

With new compounds being added to the database continuously by the international community, including new structural predictors, there is a high likelihood for an increase in accuracy over time, due to the db growing. Our method was tested on Mg-intercalation electrodes as on well as Li-intercalation electrodes, then compared between the two classes of materials to identify correlations on the validity of our predictors.

We have a minimum of two predictors per property of the material at any given run. This is due to how we defined each battery. They have one charged and one discharged state, and only one target, at any given run. For any feature we have one value for the charged material, and one for the discharged material, as a minimum. This means that our charged- and discharged half cell configurations (X) predict any battery target property (y).

As an example, the battery $\mathrm{Mg(CrS_2)_2}$ with the battery ID: $mvc-1200000091$, presents two material ID's, one for the discharged state, $\mathrm{Mg(CrS_2)_2}$ - ($mvc-91$), and one for the charged state, $\mathrm{CrS_2}$, - ($mvc-14769$).

4.1.1 Scaling of database

The sum of features that uniquely represent each compound in the Li data set are 238 and 177 for the Mg data set. The feature vectors are vastly different and their numerical values range from thousands down to small fractions. Therefore they were normalized to improve the prediction capability, for faster training of the model and to avoid any bias preference for a particular feature. The inputs were normalized by Scikit-learn's StandardScaler, removing the mean and scaling to unit variance, so that our data look like normally distributed. The standard score, z, is computed as:

$$z = \frac{(x - u)}{std} \tag{4.1}$$

where std is the standard deviation, x is our data, and u is the mean. Our data are then centered and scaled. This is done while training our model, while the target values y where not normalized. PCA was run as explained in section 3.5.

4.2 Volumetric number density

The number of atoms and atom types in the unit cell is an extensive quantity, *i.e.* it depends on the dimensions of the cell. For this reason it should not be directly provided as a descriptor. To make it a intensive quantity (independent of the size of the system), we dived the number of atom types with the volume of the unit cell. This is the volumetric number density, *vnd*, is used to describe concentration of countable objects. It is defined as:

$$n = \frac{\#\text{ of atoms}}{\text{Volume}} \tag{4.2}$$

Where Volume is the volume of the unit cell.

In the vnd, there is a predictor for each individual element. That is; if the intercalation battery is $Mg(CrS_2)_2$ then the the number densities for; magnesium, chromium, and sulfur, related only to that material, will be predictors. The charged material, CrS_2 will have the predictors with values: $S_{vol} = 36.6292$ and $Cr_{vol} = 18.3150$. Instead the discharged material $Mg(CrS_2)_2$ will have the vnd-predictors: $S_{vol-dis} = 30.1286$, $Cr_{vol-dis} = 18.3150$, and $Mg_{vol-dis} = 7.5321$. For the Mg-ion db, and the Li-ion db, a total of 30 and 53 different types of atoms were identified. All other elements still exist as predictors for this framework, both charged and discharged, and exist as possible branches in the decision threes, but they are given the value 0, unless they share the same elements. This quality is unique for the volumetric number density, All other predictors minimize "empty" features.

It is probable that such a direct measurement of a geometrical aspect could be a good predictor due to the possibility of finding correlation between our targets and the type and intensity of atoms in a given cell. The reason for not giving our algorithm the entire CIF file, is that it probably would find little correlation, due to the bias-Variance-trade-off as mentioned in section 3.2, any relevant information would be drowned in useless information, there is such a difference between some of these files that they would be difficult to normalize.

4.3 Void Fraction

Void Fraction, or the porosity, is a measurement of the void space in the material. Calculations are based on classical force fields that describe interatomic interactions between the atoms of the material and a helium atom. For the interactions of the He atom with the guest atoms, usually a Lennard Jones potential is applied. We measure the accessible void, that is, the total amount of free space in the material. It is a quantity that can be measured experimentally, and therefore has a physical meaning.

The pore volume is obtainable experimentally under the assumption that Gurvich rule is valid. It assumes that the density of the saturated nitrogen in the pores is equal to its liquid density, independent of the internal void network. The pore volume (v_{pore}) and the porosity (θ) can be computed from:

$$v_{pore} = \frac{n_{N_2}^{ads,satd}}{\rho_{N_2}^{liq}} \tag{4.3}$$

$$\theta = v_{pore} \cdot \rho_{cryst} \tag{4.4}$$

Where $n_{N_2}^{ads,satd}$ is the specific amount of nitrogen adsorbed, $\rho_{N_2}^{liq}$ is the density of liquid nitrogen, and ρ_{cryst} is the density of the crystal in question. It is important to understand that these experimental values does not account for all small voids between atoms where the nitrogen does not fit. There also exist pores that are non accessible for a nitrogen atom, so the experimental data using nitrogen is different from the DFT calculations using *e.g.* helium.

There are many different computational methods to obtain the pore volume. Where each one compute slightly different portions of the full volume, for this thesis we used the computational structure characterization tool Poreblazer [90]. We have used two different approaches for measuring the pore volume. First, the geometric pore volume, Ge_{pv}, which is defined as all the free volume of the unit cell. Secondly, helium pore volume, He_{pv}. Helium pore volume does not use hard-sphere interactions between the probe atom and the atoms of material, instead it uses a more realistic intermolecular potential. This makes the calculations temperature

dependent. The calculation are done at $298\,\mathrm{K}$. More details on these calculations can be found in the file default.dat or the supportive information given in github [**github**].

Void Fraction originates as a characterization method for microporous crystals and have had great success in metal organic frameworks (MOFS), as demonstrated also by our group of collaborators [20], [21].

Void fraction calculations used the molecular mechanics force field, Universal force field (UFF), database from "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations" [91]. Wherein the force field parameters are estimated using general rules based on the elements, their hybridization, and their connectivity.

4.4 AP-RDF Descriptors of Electrode materials

Atomic property weighted radial distribution function (AP-RDF) was found to be a good predictor by Fernandez *et al.*, yielding R^2 values in the range from 0.70 to 0.82 as a predictor on MOFs. The authors applied PCA (3.5) and found that AP-RDF exhibited good discrimination of geometrical, and other properties [22].

The radial Distribution Function (RDF) is the interatomic separation histogram representing the weighted probability of finding a pair of atoms separated by a given distance. In a crystalline solid, the RDF plot has an infinite number of sharp peaks where the separation and height are characteristic of the lattice structure. We used the minimum image convention (boundary condition) and the RDF scores will be uniquely defined inside of the unit cell, per material-ID [92]. The RDF can be expressed as:

$$RDF^P(R) = f \sum_{i,j}^{\text{all atom pairs}} P_i P_j e^{-B(r_{ij}-R)^2} \tag{4.5}$$

In our case the RDF scores in a electrode framework has been interpreted as the weighted probability distribution to find a atom pair in a spherical volume of

radius R inside the unit cell according to equation 4.5l.

The sum is done over all the atom pairs, where R_{ij} is the minimum image convention distance of these pairs, B is a smoothing parameter, and F is a scaling or normalization factor. Our Own approach to this is written in Fortran, and can be found on Github AP-RDF, with an operational pdf.

The RDF can be weighted to fit the requirements of the chemical information to be represented, by introducing the atomic properties, P_i and P_j. We weighted the radial probabilities by three tabulated atomic properties namely electronegativity, polarizability, and Van der Waals volume, which gives us the AP-RDF. While a regular RDF function encodes geometric features, the atomic property weighted RDF additionally characterizes the chemical features within a material. An atomic property weighted RDF can be seen on the screen.

To test our method, we used it to reproduce the results for the two MOFS, namely *IRMOF-1* and *MIL-45* found in the article by Fernandez.[22]. We confirmed their findings.

We explored two approaches, the first one trying to minimize the number of new features by performing a cross-product on our data between the AP-RDF data and the original db. The original one row per battery was changes, every battery would get between 8 and 52 duplicated rows, where only 7 new features were introduced, these are: electronegativity, van der waals volume and polarization, for both charged and discharged materials. The radius was also introduced as a feature, but disregarded quickly by PCA. This was tried so that we minimize the risk of the algorithm missing columns, and rather let the amount of data that lies in proximity of each other weigh heavier.

Discussion: drowning the data, to many target values m.m.

The second approach is equal to Fernandez *et al.* approach. It creates one new feature per radius, *i.e.* 106 new features for electronegativity alone with r from 2 to 15 with a step size of 0.25 .

Part IV

Results & Discussion

5 Results & Discussion

The accuracy of the ML method in this work is evaluated using R-squared (R^2), "Mean" of the training target values, standard deviation (stdev) of the error, mean absolute error (MAE), root-mean-square-error (RMSE) and weighted absolute percentages error (WAPE), as described in 3.4. Lower RMSE, MAE, and WAPE, with a high R^2-score indicate a better ML prediction. The overall performance of the model was evaluated using the k-fold cross-validation technique for 10 folds 3.4.8, and is referred to as R^2 - cross-validation mean or R^2-CVM. Before each run a PCA is performed and the PCs that collectively accounted for 99% of the variation are selected for the RF run, the last 1% of the variation, i.e. the other predictors, are removed.

For each intercalation battery type, both Mg-ion and Li-ion, four different sets of descriptors are reported, namely:

- Material specific properties
- Volumetric number density
- Void fraction
- APRDF

The targets the suggested features are tested on are: **Average Voltage** (AV), **Gravimetric Capacity** (GC), **Volumetric Capacity** (VC), **Specific Energy** (SE), [redacted] and **Energy Density** (ED). First, the set of descriptors were individually tested. Afterwards, they were combined to make the best possible classifier.

5.1 Target distribution

From the target distributions in figure 12 and 13, we can see that the Li-ion db has more data, but also stronger outliers. An example of outlier from the Li-ion db, there are 11 batteries with an AV between $6 - 9\,\mathrm{V}$ in the db, or the single

material with a specific energy between $2500\&3000\ \mathrm{Wh/kg}$. Except for the outliers are the distribution of AV even for both databases, with most values in the range of $1-5$. This trend, of similar distribution seems valid for SE and ED, which seems curious due to Li metal having a higher SE. While GC and VC seems to be lower for Li-ion batteries in general.

As a note on the target data, GC and VC, as well as SE and ED are different representations of the same characteristic in the battery. The reason for using both as targets is to evaluate how our ML model differs between these dataset.

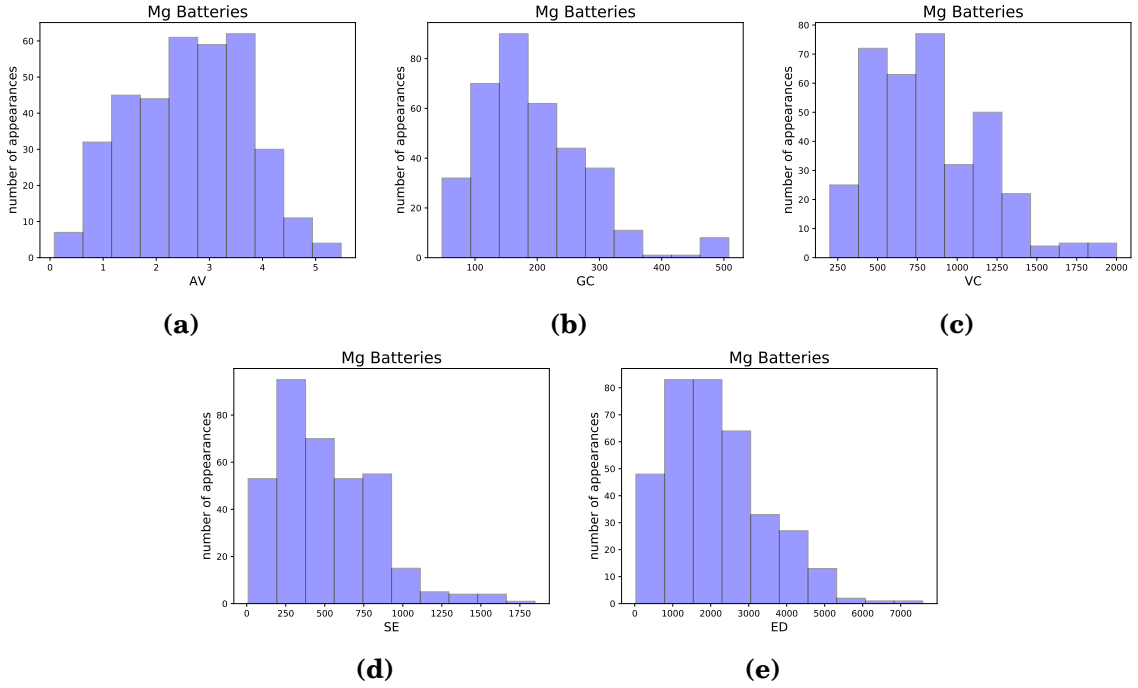


Figure 12 – a)-e) are the distribution of the targets Average Voltage, Gravimetric Capacity, Volumetric Capacity, Specific Energy and Energy Density, for the Mg-ion database. The x-axis are the target values and the y-axis are the number of appearances for that range of target values.

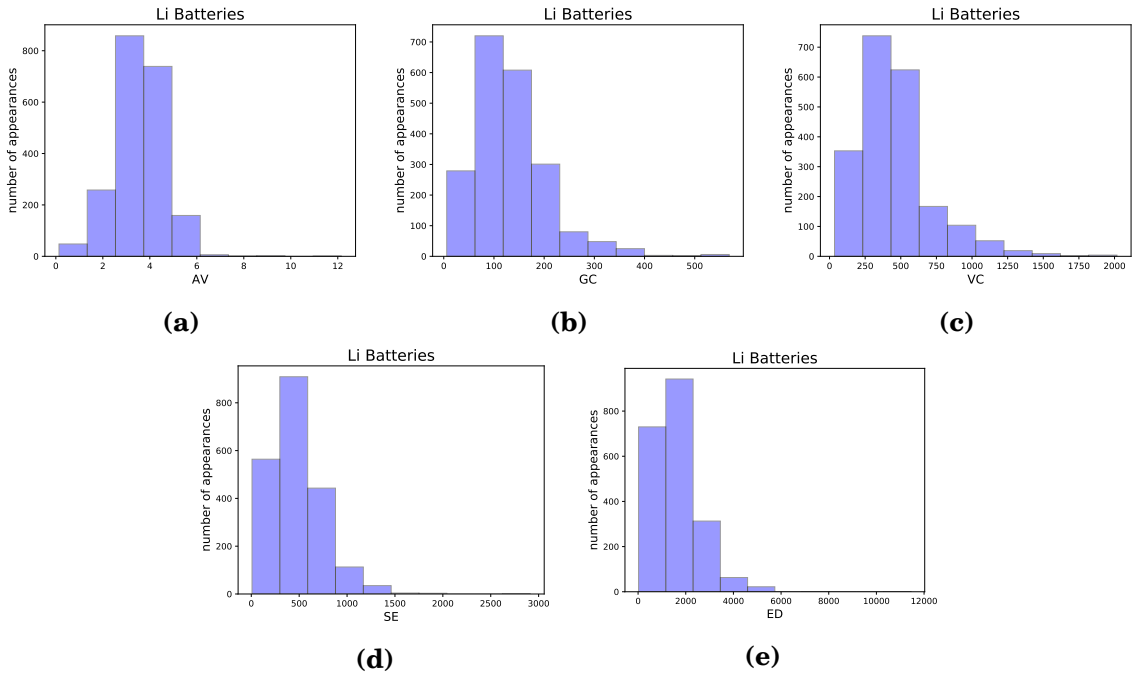
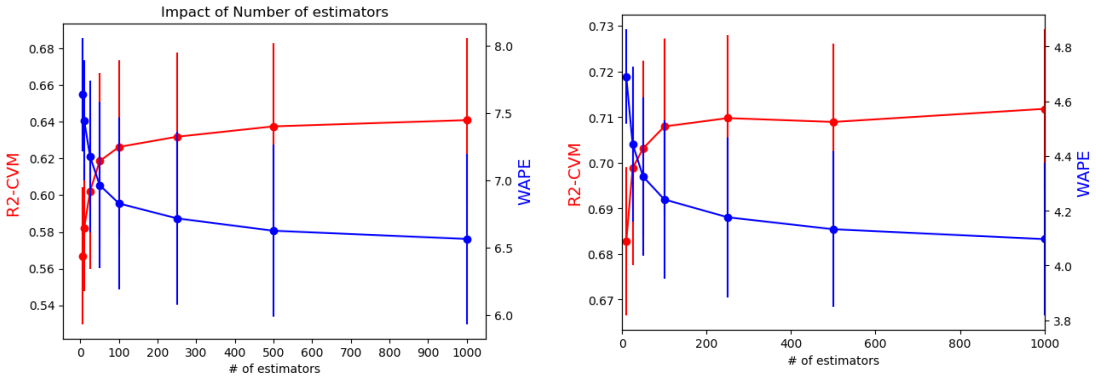


Figure 13 – a)-e) are the distribution of the targets: Average Voltage, Gravimetric Capacity, Volumetric Capacity, Specific Energy and Energy Density, for the Li-ion database. The x-axis are the target values and the y-axis are the number of appearances for that range of target values.

5.2 Size of database, and number of estimators

There are two hyper-parameters that need to be adjusted when using random forest. These are the number of estimators and the maximum number of features. The number of estimators are the sum of decision trees used in the forest. The optimal amount of trees used for the model is a compromise between predicational quality and computational cost: The more trees, the better the prediction, but the model becomes more computationally expensive. However, the improvement rate will slow down after a critical number of trees. To find the optimal number of trees for our predictions, we made several model using $5, 10, 25, 50, 100, 250, 500, 1000$ decision trees, creating 10 unique forests for each amount of trees and plottet them with the standard deviation of the 10 runs. The results, for both databases can be seen in figure 15. We decided to use 250 estimators, this gave a good compromise between cost and accuracy. It could be argued that applying 100 estimators is sufficient for an accurate prediction, but the computational impact is not noticeable for a database of this size, so opting in to a possible improvement is favored. For the

latter, the number of selected features can influence the generalization error by: Applying few features, reducing the number of trees to lower correlation among the trees, leading to a stronger forest. Or selecting many features, increasing the strength of each individual tree. In Breiman's paper on RF [65] it is recommended to apply $N/3$ features per split, with N features. While in the paper "Extremely randomized trees" by Geurts *et al.* it was empirically justified to used all features [93]. Our random forest regressor always considers all features instead of a random subset of features, due to Geurts paper, and the paper by Shandiz *et al.* [68] showing to higher accuracy by applying all features.



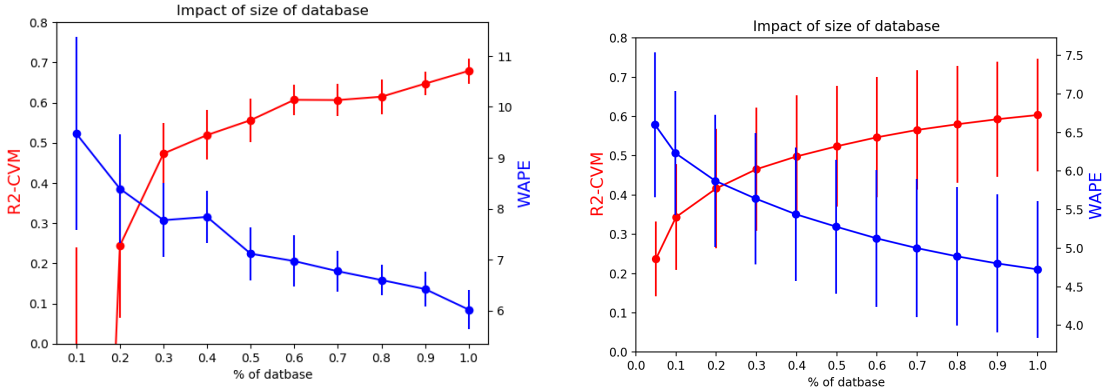
(a) Mg **(b)** Li

Figure 14 – The number of estimators used plottet vs. R^2-CVM and WAPE. The predictions are done with the same combination of features as in table 17, with Average Voltage as the target.

The two data sets are of different size. The Mg-ion database (db) has 355 different batteries, while the Li-ion db has 2073 different Li batteries, more than five times the size of the Mg-ion db. Both of these databases are relative small from a ML perspective, but are large enough to show correlations.

Impact of the size of our dataset was tested using $10\%, 20\%, ..., 100\%$ of our database. Our data were split into two parts, the training data $X_{\text{train}}, 80\%$, and the test data, $X_{\text{test}}, 20\%$. Each of these calculations were done 10 times for all unique percentage. The mean of R^2-CVM and WAPE was plottet with their respective standard deviations against the percentage of the database used, as seen in figure 15. The Li-ion db was also tested for 5% of the db, utilizing 103 Li-ion batteries, almost the same as 30% of the Mg-ion db. This was not feasible for the Mg-ion db, since 5% makes up 17 batteries, and it makes the ML model very unreliable, as

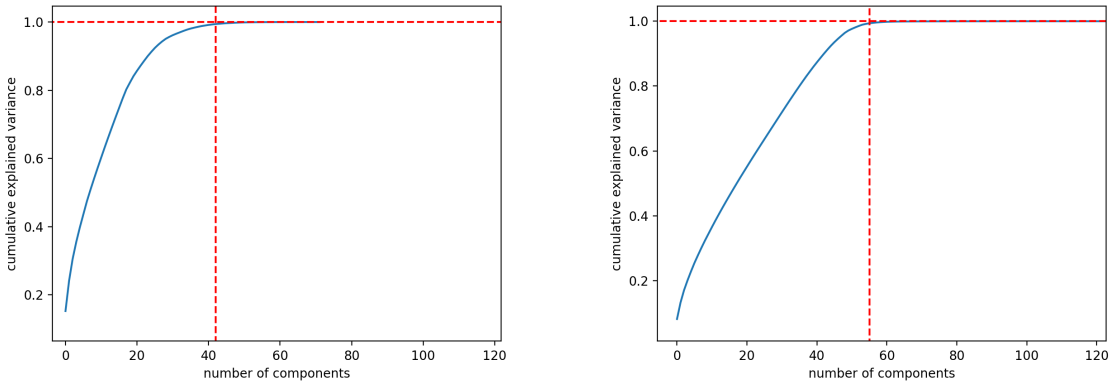
can be seen from the 10%, and 20% in figure 15.



(a) Mg **(b)** Li

Figure 15 – Indicates the impact of the size of the database. Percentage of the database is plottet vs. $R^2 - CVM$ and WAPE. The predictions are done with the same combination of features as in table 17, with Average Voltage as the target.

An example of a PCA run on our data, as explained in section 3.5, are shown in figure16. We found that we could reduce the dimensionality of our two datasets (figure 16), when applying all features, with 38.9% and 57.0% for the Mg-ion and Li-ion databases respectively, that is a reduction of 29 out of 72, and 70 out of 124 applied features.



(a) Mg **(b)** Li

Figure 16 – Principal component analysis for the feature vectors leads to a reduction of: a) 40.2% of the dimensionality, removing 29 features from the 72 predictions applied to the Mg-ion database. And b) 56.0% of the dimensionality, removing 70 of the 124 features form the prediction applied to the Li-ion database.

5.3 Material specific properties

The *material specific properties* (msp), are the properties related to every individual material, both charged and discharged. They are; **energy, energy per atom, volume of the unit cell, band gap, density, magnetization, number of sites**, and **elasticity**, for both the charged and discharged material, as explained in section 2.6. Evaluation of **stability** (stab) from the MP db will be included in this section, but will be treated as its own group of features. The Materials Project database included **formation energy per atom** for most structures, but due to "None" values and no significant change in our predictions. It is likely that any given decision tree with variables containing zeros will be ignored by the algorithm unless they facilitate some other correlation. Therefor it was decided to leave these features out of our predictions. We will first have a look at the distribution of our prediction.

Figure 17 shows the distribution for the msp, for charged materials in green and discharged materials in red. The two battery materials tends to have the same type of distribution and therefore it is interesting to see that one of the two features are commonly not dismissed, it is rather the feature as a whole that is deemed dispensable by the ML model.

The results from the Mg-ion db, table 1, shows that msp are correlated to the targets, with especially high values for AV (60.9%). The accuracy for GC,VC, SE and ED are between $47.1-56.8\%$, showing correlation for all targets and increasing error with the WAPE going from 7.9 to 13.26%. The WAPE shows that the uncertainty is larger for SE and ED than for capacity and voltage. The $6-7$ features ignored by PCA are elasticity, number of site, energy per atom, and the "charged volume".

The results from the Mg-ion db, table 1, shows that msp are correlated to the targets, with especially high values for AV (60.9%). The accuracy for GC,VC, SE and ED are between $47.1-56.8\%$, showing correlation for all targets and increasing error with the WAPE going from 7.9 to 13.26%. The WAPE shows that the uncertainty is larger for SE and ED than for capacity and voltage. The $6-7$ features ignored by PCA are elasticity, number of site, energy per atom, and the "charged volume".

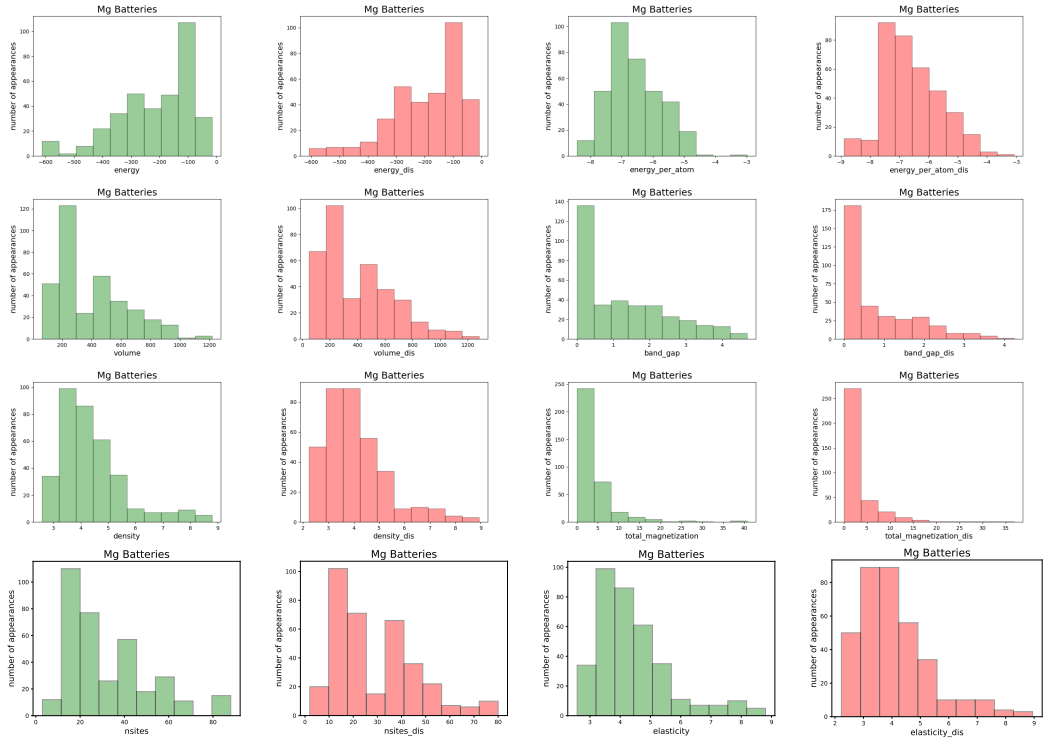


Figure 17 – These are the distributions of the 8 material specific properties that account for the most variance, for both the charged and discharged material for the Mg-ion database. The x-axis are the feature values and the y-axis are the number of appearances for that range of feature values.

R^2-CVM on AV in the Li-ion db is lower than in the Mg-ion db, but the accuracy of the predictions are higher. GC and VC have a better score, but a larger error, while SE and ED have lower errors, and smaller scores. The components prioritized are the same as for the Mg-ion db, except for a couple of runs where it removes "charged density" as a predictor, which, for both runs, had a low variance ratio, meaning it had little variance compared with the other features.

The increase in data seems to not effect our predictions to a large extent. R^2-train shows that our model is about as accurate on the data it has been presented.

Table 1 – msp prediction-scores for the Mg db tested against the given targets.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.5224	0.4023	0.47064	0.4997	0.4826
R^2-train	0.9293	0.9208	0.9078	0.9317	0.90721
Mean:	2.5830	194.105633	844.3028	503.0492	2105.3873
Stdev:	0.2905	22.8769	103.3817	79.9453	399.8795
RMSE:	0.2907	22.9191	103.3826	80.0723	400.4032
MAE:	0.2042	16.7839	74.45455	61.0444	279.353
WAPE:	7.9079	8.6468	8.8184	12.1348	13.2685
R^2-CVM:	0.60989	0.471065	0.4985	0.56822	0.5247
components used:	10/16	10/16	10/16	9/16	9/16

Table 2 – msp from the Li-db tested against the given targets.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.5384	0.4162	0.4870	0.3888	0.44016
R^2-train	0.9308	0.9146	0.9277	0.9189	0.9185
Mean:	3.58038	132.6524	463.5728	479.7717	1609.3506
Stdev:	0.2853	21.7816	70.5023	75.2707	272.7973
RMSE:	0.2853	21.804	70.5346	75.323	735.50848
MAE:	0.1969	15.3558	48.2923	55.3367	191.3195
WAPE:	5.5013	11.5759	10.4174	11.5339	11.8879
R^2 CVM:	0.5692	0.4456	0.5136	0.4572	0.4815
components used:	9/16	9/16	9/16	8/16	9/16

Stability

In this work we tried to predict the stability of a material, based on the same predictors as introduced, this did not work. One possible reasons might be that this "stability" is connected to each particular material, it is not a property of the battery. Utilizing a combination of both charged and discharged properties to predict a feature related to only one of the two types of material can confuse the ML algorithm. The predictions on charged and discharged stability can be found on github. In addition, we decided to include stability as a predictor to see if this upped our results. Using stability as a predictor did not increase the predictions over the threshold ($> 2\%$) that was set. Yet, the ML model did include it in its predictions, so it contains relevant information and will therefor be kept.

5.4 Volumetric number density

Mg-ion framework

Volumetric number density (vnd) as described (4.2) are shown in table 3,to 8, first for the Mg-database then for the Li-database. Due to the nature of vnd, and our database having both charged and discharged materials, it is necessary to try the three alternatives; only the charged materials, only the discharged materials and the combination of both the charged and the discharged materials.

Table 3 – Mg- db, the charged material as vnd predictors. A total of 21 components are applicable. Mg-prediction results on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row showes the number representing that type of evaluation, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.51920	0.1783	0.2513	0.5751	0.2864
R^2-train	0.94365	0.9055	0.91827	0.9194	0.9302
Mean:	2.5492	192.7605	848.22	537.5	2227.422
Stdev:	0.2604	24.9649	98.056	84.4191	335.7547
RMSE:	0.26043	24.965	98.0653	84.5840	336.2162
MAE:	0.1860	51.9769	70.8905	63.5310	252.03396
WAPE:	7.2986	9.2791	8.3575	11.8200	11.3150
R^2 CVM:	0.58126	0.2464	0.3964	0.5280	0.5174
components :	19/21	19/21	19/21	19/21	19/21

There are a couple of different results that are particularly interesting. First of all; The use of all vnd-predictors yield the best predictions in all cases, with the lowest error in almost all. When combining the two group of predictors PCA tells us that there is an overlap in the information from the charged and discharged materials as given by the number of components used by the ML algorithm. *i.e.* for charged materials $19/21$ components where used to account for 99% of the variance in the data, for the discharged materials $22/30$ components where used, while when combined 30 out of the total 51 components were necessary to achieve the same 99%. This shows an overlap in the two group of predictors data.

There are targets that respond better to the use of one state of material, charged or discharged, than the other. Gravimetric and volumetric capacity re-

Table 4 – Mg-db, the discharged material as vnd-predictors. A total of 30 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row showes the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.4879	0.7213	0.5941	0.4612	0.4779
R^2-train	0.9312	0.92499	0.9332	0.9496	0.9218
Mean:	2.7038	184.0281	869.2605	483.4859	2288.429
Stdev:	0.2835	24.0464	86.9195	75.7915	352.595
RMSE:	0.2836	24.04749	86.9622	75.8086	352.6381
MAE:	0.1938	14.4620	44.5180	57.4124	221.0466
WAPE:	7.1708	7.8586	5.1213	11.8746	9.6593
R^2 CVM:	0.5885	0.6190	0.64017	0.6115	0.5980
components:	22/30	22/30	22/30	22/30	23/30

sponds particularly well to the discharged materials with a R^2-CVM $= 0.6226$, but improve (by a small percentage $3-5\%$) when given the combination of materials, and the error decreases. AV is the only target that has the same prediction for both charged and discharged, but it is a clear improvement when combining both with an increase in R^2-CVM of around 4%. This seems reasonable due to the discharged materials containing more information because of the discharged material set having a larger number of atom types, and that atom (Mg) being the only atom all batteries have in common (i.e. the battery framework $\mathrm{Mg}_{0-1}\mathrm{CrF}_6$ with CrF_6 being the charged material and MgCrF_6 being the discharged material).

When combining the two group of predictors our predictions on all five targets are between 62 to 68 percent, with high certainty. Which showes that this is a good predictor for all of these targets. Noticeably the R^2-train is also considerably better both for the combination of materials and for the msp.

Lastly, the quality of our estimator is significantly lower for specific energy and energy density. It seems that the quality of our predictor drops on these predictors. (SP $-\mathrm{Wh/kg}$, ED $-\mathrm{Wh/l}$). why? whywhywhy?

Table 5 – Mg-db using both the discharged- and the charged materials as vnd-predictors. A total of 51 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row showes the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.6094	0.6150	0.7993	0.5685	0.6560
R^2-train	0.9507	0.9402	0.9452	0.9540	0.9385
Mean:	2.7871	194.4788	825.5211	483.6478	2166.943
Stdev:	0.2405	20.3516	80.1226	69.5891	314.8012
RMSE:	0.2409	20.3608	80.1511	69.59	314.8037
MAE:	0.1857	11.9768	38.6913	52.7039	219.9826
WAPE:	6.6656	6.1584	4.6868	10.8971	10.1517
R^2 CVM:	0.6261	0.6622	0.6758	0.64942	0.6532
components:	30/51	27/51	29/51	29/51	27/51

Predictions on Li-ion intercalation frameworks with vnd

The same technique, as used in the previous section, was applied to the Li-ion db, with 2108 battery frameworks instead of 365, as were the case for Mg-ion intercalation type db. The Li-ion db is a bigger database with different and more unique atom types, which is likely to calls for more variability, possibly more bias, and most likely a lower uncertainty in the predictions.

First of all, it is no obvious improvement when using only the discharged materials, over only the charged materials, as were with the Mg-ion db.

If we combining the two, we quickly see that the variation in the data behaves in the same way as for the Mg db, there is an obvious overlap in information and correlation in the predictors. Volumetric capacity, with the overall best predictions (71.86%), only uses 38 components, out of the possible 107 (to account for 99% of the variance). When using the charged and discharged predictors the machine learning algorithm used $38/54$ components for the charged material and $34/53$ components for the discharged components. It stands to reason that volumetric capacity would be a good target for vnd due to the intrinsic relation to volume. Because of this, it stands to reason that it should be a better predictor for ED than for SE, as it is, with an increase of 9%. The features that PCA removes are the atom types

with rare occurrences, these features mostly contain zeroes, and if they contain other values they are hard to make relevant decision rules from.

Average voltage predictions are 56% with high certainty, lower than the predictions for the Mg-ion db, but a bit higher accuracy. The AV lacks a obvious connection to the atom types in them selvs and volume. Therefore one can hypothesis that it will get better when combined with a representation of energy from the msp or electronegativity through AP-RDF.

Table 6 – Li - vnd charged.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.4466	0.22431	0.4248	0.40232	0.3402
R^2-train	0.9321	0.9060	0.9170	0.9166	0.9287
Mean:	3.589	134.03	467.40	480.46	1628.3
Stdev:	0.2884	22.165	73.304	75.481	241.6
RMSE:	0.2888	22.174	73.313	75.538	241.8
MAE:	0.1952	15.371	49.472	53.076	176.38
WAPE:	5.440	11.469	10.584	11.047	10.833
R^2 CVM:	0.5437	0.3692	0.4352	0.4324	0.4690
components used:	37/54	41/54	38/54	38/54	35/54

Table 7 – Li- vnd discharged.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.5099	0.3731	0.4539	0.3133	0.3519
R^2-train	0.9269	0.9163	0.9248	0.9178	0.9122
Mean:	3.5455	133.12	476.6359	473.7458	1609.5469
Stdev:	0.2858	21.237	70.9524	76.4492	292.0871
RMSE:	0.2860	21.2419	70.95331	76.4636	292.1071
MAE:	0.1895	14.1890	45.2840	54.9314	177.8518
WAPE:	5.3454	10.659	9.501	11.5951	11.0498
R^2 CVM:	0.5319	0.3983	0.4724	0.4367	0.44020
components used:	40/53	38/53	34/53	39/53	39/53

Table 8 – Li- vnd both charged and discharged

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.4755	0.54190	0.6923	0.5388	0.5912
R^2-train	0.9355	0.94517	0.9550	0.9351	0.9476
Mean:	3.616	135.22	460.05	473.75	1643.15
Stdev:	0.2657	17.383	54.816	69.721	213.12
RMSE:	0.2661	17.390	54.830	69.779	213.17
MAE:	0.1859	11.089	32.248	46.392	144.95
WAPE:	5.141	8.201	7.0095	9.793	8.822
R^2 CVM:	0.5608	0.6191	0.7186	0.5619	0.6506
components used:	39/107	37/107	38/107	38/107	40/107

5.5 Void fraction

Results for predictions using only the **void fraction** (vf) methods predictors, that is, the charged and discharged materials, helium volume and geometric volume. First on the Mg-ion db, then on the Li-ion db.

On its role as a predictor, the void fraction could be a good predictor, both from its success on predictions of other type of materials and the changes that occur in the space occupied by the ion in the discharged material against the charged material. We will first have a look at the distribution of the features set.

5.5.1 Distribution

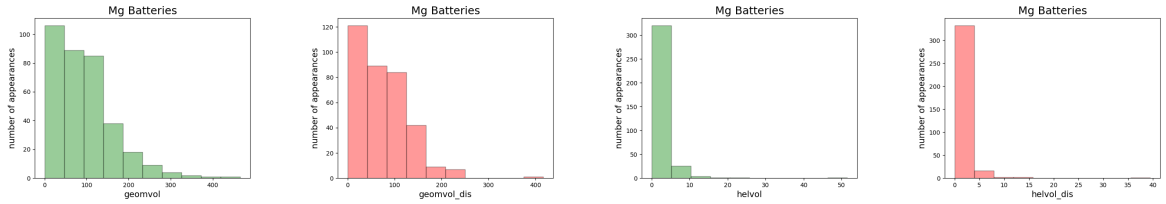


Figure 18 – Distribution of the void fraction predictors, geometric volume and helium volume, as calculated by Poreblazer for the Mg-ion db. Charged features in green and discharged features in red

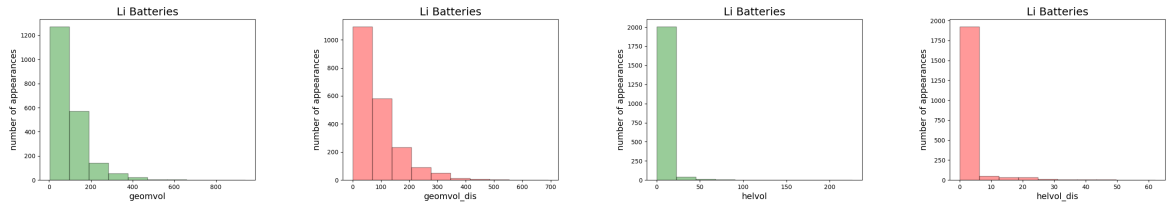


Figure 19 – Distribution of the void fraction predictors, geometric volume and helium volume, as calculated by Poreblazer for the Li-ion db.

5.5.2 Predictions

The results for vf as a predictor for the Mg-ion db is shown in table 9. Void fraction seems to be inaccurate for most targets, with the best accuracy score of 39.4% on volumetric capacity. Second best prediction is on gravimetric capacity (34.9%), which is closly related. The other predictions have to low correlation to consider them as predictors, so AV, SE and ED will be omitted for the rest of this subsection.

Table 9 – Mg- db prediction on the targets AV, GC, VC, SE, ED. A total of 4 predictors where used in this run.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	-0.3240	0.2858	0.2387	0.12553	0.06999
R^2-train	0.8621	0.9047	0.9266	0.8821	0.89021
Mean:	2.657	185.0915	830.5563	492.9507	2181.7746
Stdev:	0.4195	27.3274	93.0447	111.29816	420.6466
RMSE:	0.4197	27.3296	93.07669	111.3178	420.6466
MAE:	0.3229	19.6535	68.1477	88.04975	324.1471
WAPE:	12.1522	10.6182	8.20506	17.8617	14.85704
R^2 CVM:	0.06214	0.3490	0.3943	0.09092	0.18253
components:	3/4	4/4	4/4	3/4	3/4

On the Li-ion db (presented in table 10) the same trend as on Mg-ion db seems persistent. More data seems to emphasize that the void fraction is not a inadequate descriptor for the given targets. The best predictions are still on VC with 15% accuracy, followed by GC with 5% accuracy, but the scores being this low, and .

In all runs, three to four feature were needed to account for 99% of the variability. Between $80 - 90\%$ of the variability can be traced back to one predictor, namely geometric volume for the discharged materials, and yet it is only possible

Table 10 – Li- db prediction on the targets AV, GC, VC, SE, ED. A total of 4 predictors where used in this run.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	-0.0373	0.0603	0.05857	-0.1374	0.01655
R^2-train	0.8510	0.87095	0.89077	0.8601	0.8669
Mean:	3.5262	133.9714	447.216	457.3083	1636.7998
Stdev:	0.3914	26.2527	90.6257	102.7326	345.827
RMSE:	0.0455	26.253	90.6261	102.74241	345.8724
MAE:	0.3028	19.6381	66.8044	76.8657	261.30785
WAPE:	8.5886	14.6584	14.9378	16.8082	15.9645
R^2 CVM:	-0.06318	0.05186	0.159419	-0.02663	0.02518
components:	3/4	4/4	4/4	4/4	4/4

to get a R^2-CVM prediction above 0% for any target, by including all 4 predictors. This points at void fraction, as we have approached it here, being a bad descriptor for the given targets.

It seems reasonable that 90% of the variability are in one predictor due to these 4 predictors measuring the same physical , this physical feature seems to be expressed by having some correlation with the volumetric capacity, but this correlation might be covered by noise as we introduce more data.

Void fraction was tested as a predictor for VC, and other targets, in combination with other predictors, but due to drops in predictive capability the results are omitted from the result section. These results can be found on github.

5.6 Atomic property weighted radial distribution function

5.6.1 Row approach to AP-RDF

The second approach gave correlation with the targets as presented in table 11 for the Mg-ion db and in table 12 for the Li-ion db, here we added a new columns per value from the RDF, as explained in the section 4.4.

This approach seems promising with scores that indicates correlation for all targets, that increases on AV and ED for the Li-ion. The WAPE decreases for

AV and SE but increases for GC and VC, for a larger db witch indicates. The correlation is stable when we switch to the Li-ion db, which indicates that we have represented a property in a reasonable fashion. The second approach to AP-RDF will be discussed in more detail when the combined predictions are presented.

Table 11 – Mg- db prediction on the targets AV, GC, VC, SE, ED. A total of 106 components are applicable. A total of predictors where used in this run.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.1793	0.3855	0.3334	0.3711	0.3206
R^2-train	0.8741	0.8986	0.92503	0.88771	0.8978
Mean:	2.7054	200.669	836.204	519.0915	2250.88
Stdev:	0.4042	25.4534	89.270	103.0931	395.374
RMSE:	0.4041	25.464	89.3011	103.19	395.53
MAE:	0.3163	18.420	63.552	77.924	290.96
WAPE:	11.6913	9.1792	7.6001	15.0115	12.9264
R^2 CVM:	0.2336	0.3956	0.4438	0.3611	0.32963
components:	18/106	46/106	35/106	17/106	35/106

Table 12 – Li db prediction on the targets AV, GC, VC, SE, ED. A total of 106 components are applicable. A total of predictors where used in this run.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.2894	0.2732	0.4094	0.2666	0.3280
R^2-train	0.8951	0.9032	0.9191	0.9003	0.9043
Mean:	3.4851	134.104	460.5513	476.3011	1590.0499
Stdev:	0.3428	22.71	74.7006	85.6635	295.106
RMSE:	0.2503	22.7127	74.706	85.6724	295.1143
MAE:	0.6937	15.74474	52.9488	61.4651	202.1444
WAPE:	7.1808	11.7406	11.4968	12.90	12.7131
R^2 CVM:	0.3075	0.3379	0.4474	0.3502	0.3771
components:	42/106	46/106	44/106	46/106	46/106

5.6.2 Cross-product approach

The AP-RDF was tested with a cross-product approach, this is explained in the method section 4.4. As is clear from the tables 13 the first approach to AP-RDF did not give a good correlation with the targets and the error is to big, and will not be considered further. The idea was that it is better for a RF approach to consider longer trees rather than more features. These results are still of interest

in a ML perspective. When combined with other predictors it was clear that the ML model learned memorized what results belonged to what predictors. It could do this because of how the test data was randomly split, not accounting for target duplicates, but when tested with only the 7 features belonging to this AP-RDF approach it did not manage to memorizing the data. Pointing at it not being enough correlation in the data to find a solid pattern. The cross-product approach will not be considered further, and due to the size of such calculations the Li-ion db was also omitted from testing.

Table 13 – Mg db prediction on the targets AV, GC, VC, SE, ED. A total of 6 predictors where used in this run; radius, electronegative, van der waals volume and polarization, all for both charged and discharged materials.

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.0284	0.0799	0.1116	0.0323	0.0540
R^2-train	0.40632	0.37274	0.3877	0.3927	0.3998
Mean:	2.6294	196.4378	826.694	509.643	2120.2996
Stdev:	0.8531	69.896	279.4123	244.658	990.4840
RMSE:	1.1041	85.8334	279.413	244.6616	990.4908
MAE:	0.637	47.4149	192.349	175.8229	703.5917
WAPE:	24.2602	24.1374	23.267	34.4992	33.1835
R^2 CVM:	0.0309	0.11039	0.13321	0.04783	0.05931
components:	6/6	6/6	5/6	6/6	6/6

5.7 Combining predictors

In this subsection predictors combined. The combination of msp and vnd are of particular interest and will therefore get special attention table 14 and 15, the predictors: msp and vnd, were combined. The Mg-ion db predictions increase for AV and SE (3%, 7%), while it stays still for GC, VC and ED (change $< 2\%$). The WAPE falls which indicates that the predictions are more reliable.

The combining msp and vnd, for the Li-ion db, the trends are consistent with the Mg-ion db. For AV and SE the predictions increases (13%, and 11%) and WAPE goes down. For the predictions in general, the combination of predictors either heightens the predictive score, lowers the error or both.

Figure 20 and 21 summarizes the R^2-CVM for all targets, over the predictors

Table 14 – Mg-db applying msp and vnd. A total of 66 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row showes the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.6205	0.5940	0.66789	0.6710	0.5324
R^2-train	0.9445	0.9512	0.9611	0.94143	0.9334
Mean:	2.6786	194.662	828.9436	548.585	2142.81
Stdev:	0.2612	18.9257	67.7563	68.168	341.56
RMSE:	0.2612	18.9269	67.7746	68.1804	341.58
MAE:	0.1881	10.8578	34.5623	49.1891	219.49
WAPE:	7.0254	5.57778	4.1694	8.9665	10.2428
R^2 CVM:	0.6618	0.6661	0.6930	0.7209	0.6424
components:	34/66	35/66	35/66	33/66	34/66

presented. The utmost right labeled "combo" is the combination of the predictors msp (with stability), void fraction and vnd. "APRDF1" are not included in figure 21 as stated earlier.

Table 15 – Li-db applying msp and vnd. A total of 123 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row showes the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.5721	0.6090	0.6824	0.66525	0.6344
R^2-train	0.9609	0.9533	0.9574	0.9455	0.9528
Mean:	3.6302	131.3345	467.1239	471.7594	1664.6278
Stdev:	0.2022	15.6615	54.1556	62.4251	196.7591
RMSE:	0.20225	15.6676	54.16726	62.4668	196.8365
MAE:	0.1474	10.8578	34.45520	41.13432	135.256
WAPE:	4.0622	7.86381	7.37602	8.7193	8.1252
R^2 CVM:	0.6979	0.6444	0.71029	0.6713	0.6590
components:	43/123	44/123	45/123	46/123	45/123

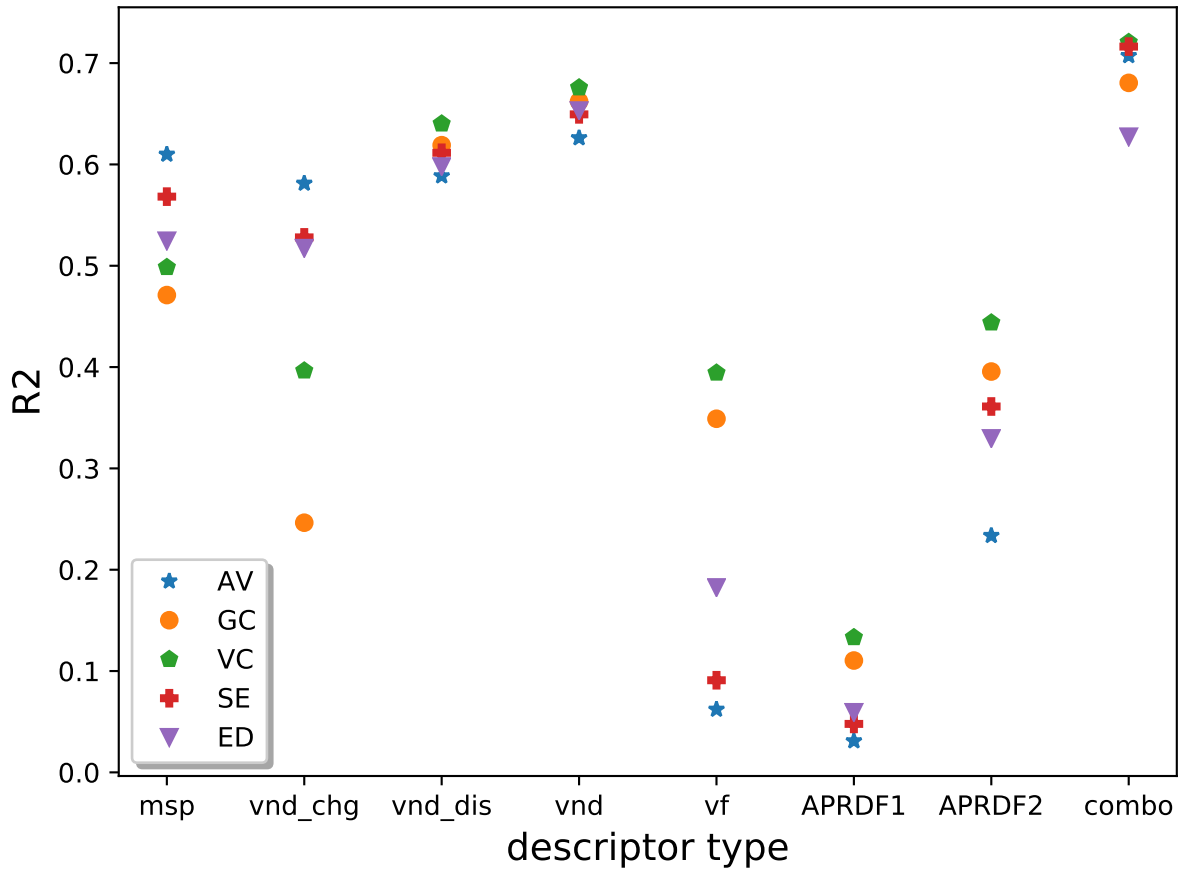


Figure 20 – Mg

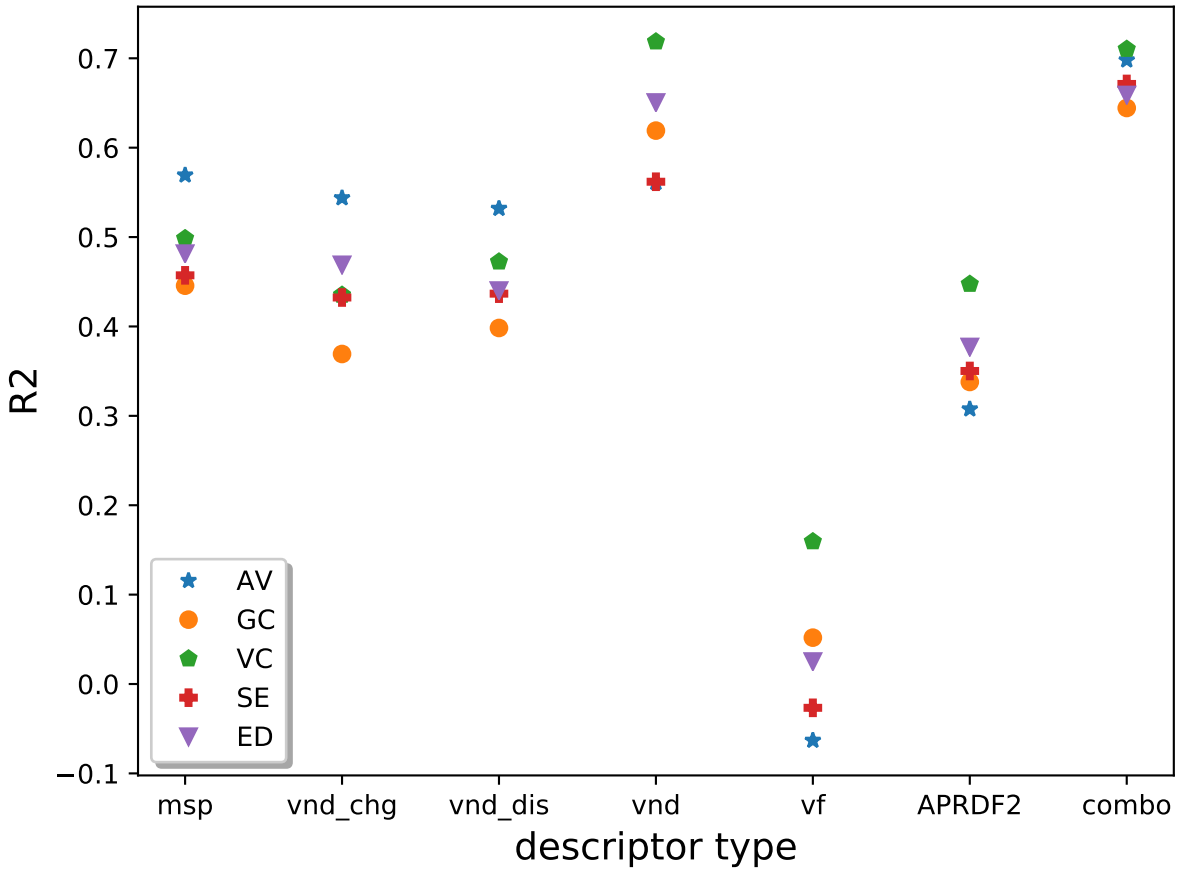

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
R2
AV
GC
VC
SE
ED
msp
vnd_chg
vnd_dis
vnd
vf
APRDF2
combo
descriptor type

Figure 21 – Li

5.7.1 Combining predictors and targets

In a last effort to evaluate the method some of the targets were introduced as predictors. The idea being is that if one of the targets is simple to compute, with *e.g.* DFT, then it is possible to predict the other targets without preforming costly calculations. In essence a way to speed up high-throughput material science. First the Mg-ion db will be considered, then the Li-ion db and at the end a combination of different targets will briefly be looked at.

Table 16 – Mg-db applying msp, vnd, stability and void fraction. A total of 72 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row shows the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.6748	0.6743	0.6952	0.6786	0.5836
R^2-train	0.9477	0.9519	0.9618	0.9580	0.9515
Mean:	2.6624	193.6127	826.9859	516.2887	2188.451
Stdev:	0.2406	18.3104	68.3938	64.8012	280.0648
RMSE:	0.2406	18.3146	68.4006	64.8012	280.0846
MAE:	0.1719	10.9457	34.9866	47.5026	192.823906
WAPE:	6.4566	5.65344	4.2306	9.2008	8.81097
R^2 CVM:	0.7072	0.6805	0.7206	0.7163	0.62737
components:	39/72	35/72	37/72	34/72	39/172

Table 17 – Li-db applying msp, vnd, stability and void fraction. A total of 131 components are applicable. Predictions on the targets; Average Voltage (AV), gravimetric capacity (GV), volumetric capacity (VC), specific energy (SE), and energy density (ED). Each row shows the number representing that type of test, as included in section (3.4).

$\frac{Target:\rightarrow}{Accuracy:\downarrow}$	AV	GC	VC	SE	ED
R^2-score	0.6765	0.6367	0.67698	0.6657	0.6368
R^2-train	0.9580	0.9491	0.9575	0.9443	0.9497
Mean:	3.5426	134.2437	460.6255	472.3159	1598.6273
Stdev:	0.2164	16.62927	54.9892	64.09351	214.561
RMSE:	0.2164	16.6336	55.0011	64.1256	214.5891
MAE:	0.14963	10.6198	34.9866	42.26381	136.5968
WAPE:	4.2236	7.91080	7.70540	8.9482	8.5446
R^2 CVM:	0.7317	0.6472	0.7151	0.66435	0.6861
components:	47/131	47/131	49/131	51/131	45/131

Finally, the first half of figure 22 shows how the R^2-CVM increases when com-

bining predictors. msp are first plotted alone, before stability, volumetric number density and void fraction is added. The second part, starting at "co" (the predictors that far), includes some of the targets as the predictor. The targets are in the x-axis denoted AV for the voltage, GCVC for Gravimetric Capacity and Volumetric Capacity, and SEED for Specific Energy and Energy Density.

the capacity or specific energy are used as a predictor for the other targets the predictions are hovering above 97% for all targets but AV. This shows that it is possible to get high enough predictions, given the right predictors, and that some vital component is lacking for the ML algorithm to achieve such an estimate. For the target average voltage, some property that is not in the capacity or specific energy, or any of the other predictors is missing.

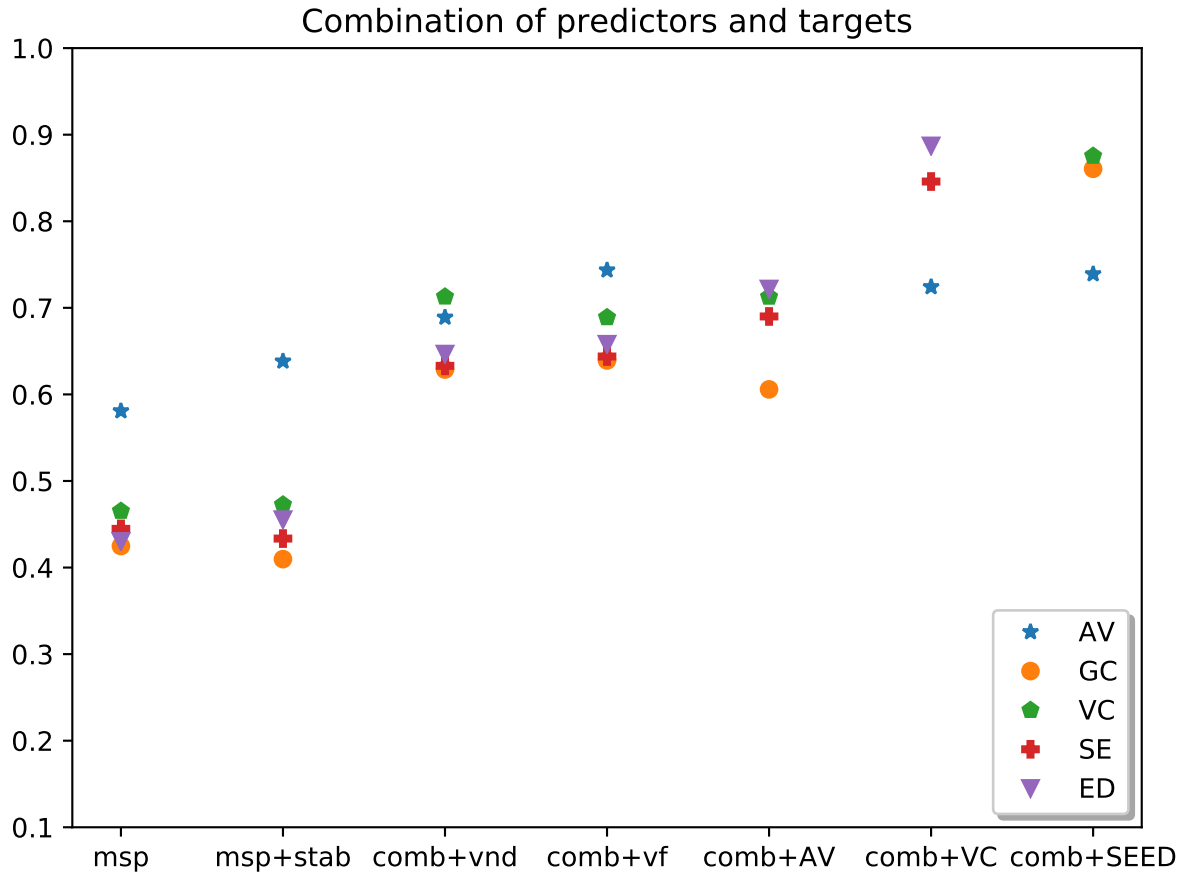


Figure 22 – Li

For *average voltage* ur best accuracy is around 73% 17, with an high accuracy, WAPE $= 4.2\%$. Which is not adequate for our ML model to be reliable on prediction

of AV. The paper by Joshi *et al.* [81] made predictions on the same db, where they applied the features: the concentration of the active metal ion, crystal lattice types and space group numbers. They also adopted features from the work by Ward *et al.* [**ward2016general**], where the elemental properties where added to the feature vector.

Part V

Summary

6 Conclusion and future work

6.1 Batteries

6.2 future work

6.2.1 improving method

References

[1] George E Blomgren. “The development and future of lithium ion batteries”. In: *Journal of The Electrochemical Society* 164.1 (2016), A5019.

[2] Tomooki Hosaka et al. “Research Development on K-Ion Batteries”. In: *Chemical Reviews* (2020).

[3] Wenzhuo Cao, Jienan Zhang, and Hong Li. “Batteries with high theoretical energy densities”. In: *Energy Storage Materials* 26 (2020), pp. 46–55.

[4] Yaokun Pang et al. “Additive Manufacturing of Batteries”. In: *Advanced Functional Materials* 30.1 (2020), p. 1906244.

[5] International Energy Agency. “Global EV Outlook 2018: Towards Cross-modal Electrification”. In: IEA. 2018.

[6] Rui Xiong. “Overview of Battery and Its Management”. In: *Battery Management Algorithm for Electric Vehicles*. Springer, 2020, pp. 1–24.

[7] Gabriel R Schleder et al. “From DFT to machine learning: recent approaches to materials science–a review”. In: *Journal of Physics: Materials* 2.3 (2019), p. 032001.

[8] Nir Pour et al. “Structural analysis of electrolyte solutions for rechargeable Mg batteries by stereoscopic means and DFT calculations”. In: *Journal of the American Chemical Society* 133.16 (2011), pp. 6270–6278.

[9] Scott Kirklin, Bryce Meredig, and Chris Wolverton. “High-throughput computational screening of new Li-ion battery anode materials”. In: *Advanced Energy Materials* 3.2 (2013), pp. 252–262.

[10] Mayumi Kaneko et al. “Local Structural Studies of LiCr y Mn2-y O4 Cathode Materials for Li-Ion Batteries”. In: *The Journal of Physical Chemistry B* 107.8 (2003), pp. 1727–1733.

[11] Brett Ammundsen et al. “Lattice dynamics and vibrational spectra of lithium manganese oxides: A computer simulation and spectroscopic study”. In: *The Journal of Physical Chemistry B* 103.25 (1999), pp. 5175–5180.

[12] MS Islam and B Ammundsen. “Atomistic Computer Modelling of Oxide Cathode Materials for Lithium Ion Batteries”. In: *Materials for Lithium-Ion Batteries*. Springer, 2000, pp. 293–307.

[13] J Spencer Braithwaite et al. “A computational study of the high voltage Li x Co y Mn 4- y O 8 cathode material”. In: *Physical Chemistry Chemical Physics* 2.17 (2000), pp. 3841–3846.

[14] Stefano Curtarolo et al. “The high-throughput highway to computational materials design”. In: *Nature materials* 12.3 (2013), pp. 191–201.

[15] Anubhav Jain et al. “The Materials Project: A materials genome approach to accelerating materials innovation”. In: *APL Materials* 1.1 (2013), p. 011002. ISSN: 2166532X. DOI: 10.1063/1.4812323. URL: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1%5C&Agg=doi.

[16] Stefano Curtarolo et al. “AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations”. In: *Computational Materials Science* 58 (2012), pp. 227–235.

[17] James E Saal et al. “Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)”. In: *Jom* 65.11 (2013), pp. 1501–1509.

[18] Claudia Draxl and Matthias Scheffler. “NOMAD: The FAIR concept for big data-driven materials science”. In: *MRS Bulletin* 43.9 (2018), pp. 676–682.

[19] Austin D Sendek et al. “Machine learning-assisted discovery of solid Li-Ion conducting materials”. In: *Chemistry of Materials* 31.2 (2018), pp. 342–352.

[20] George S Fanourgakis et al. “A Robust Machine Learning Algorithm for the Prediction of Methane Adsorption in Nanoporous Materials”. In: *The Journal of Physical Chemistry A* 123.28 (2019), pp. 6080–6087.

[21] George S Fanourgakis et al. “A Universal Machine Learning Algorithm for Large Scale Screening of Materials”. In: *Journal of the American Chemical Society* (2020).

[22] Michael Fernandez, Nicholas R Trefiak, and Tom K Woo. “Atomic property weighted radial distribution functions descriptors of metal–organic frameworks for the prediction of gas uptake capacity”. In: *The Journal of Physical Chemistry C* 117.27 (2013), pp. 14095–14105.

[23] Alexander Volta and Joseph Banks. “I. On the electricity excited by the mere contact of conducting substances of different kinds”. In: *The Philosophical Magazine* 7.28 (1800), pp. 289–311.

[24] Franco Decker. “Volta and the pile”. In: *Electrochemistry Encyclopedia* (2005).

[25] John Frederic Daniell. “XI. Additional observations on voltaic combinations. In a letter addressed to Michael Faraday, DCLFRS, Fullerian Prof. Chem. Royal Institution, Corr. Memb. Royal & Imp. Acadd. of Science, Paris, Petersburgh, &c. By J. Frederic Daniell, FRS, Prof. Chem. in King’s College, London”. In: *Philosophical Transactions of the Royal Society of London* 126 (1836), pp. 125–129.

[26] Wikipedia. *Wikipedia, The Free Encyclopedia*. [Online; accessed 25-February-2020]. 2004. URL: %5Curl%7Bhttps://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Galvanic_cell_labeled.svg/600px-Galvanic_cell_labeled.svg.png%7D.

[27] Claus Daniel and Jürgen O Besenhard. *Handbook of battery materials*. John Wiley & Sons, 2012.

[28] M Stanley Whittingham and Fred R Gamble Jr. “The lithium intercalates of the transition metal dichalcogenides”. In: *Materials Research Bulletin* 10.5 (1975), pp. 363–371.

[29] Materials Project. *Materials Project TiS2, High-throughput Identification and Characterization of Two-dimensional Materials using Density functional theory*. [Online; accessed 26-February-2020]. 2020. URL: %5Curl%7B%20https://materialsproject.org/materials/mp-558110/%7D.

[30] Nobel Media AB 2020. *The Nobel Prize in Chemistry 2019*. [Online; accessed 11-March-2020]. 2020. URL: %5Curl%7Bhttps://www.nobelprize.org/prizes/chemistry/2019/summary/%7D.

[31] K Mizushima et al. “LixCoO2 (0< x<-1): A new cathode material for batteries of high energy density”. In: *Materials Research Bulletin* 15.6 (1980), pp. 783–789.

[32] John B Goodenough, K Mizushima, and T Takeda. “Solid-solution oxides for storage-battery electrodes”. In: *Japanese Journal of Applied Physics* 19.S3 (1980), p. 305.

[33] John B Goodenough and Kyu-Sung Park. “The Li-ion rechargeable battery: a perspective”. In: *Journal of the American Chemical Society* 135.4 (2013), pp. 1167–1176.

[34] Wikipedia. *Wikipedia, The Free Encyclopedia*. [Online; accessed 26-February-2020]. 2004. URL: %5Curl%7Bhttps://upload.wikimedia.org/wikipedia/commons/5/57/Lithium-cobalt-oxide-3D-balls.png%7D.

[35] Tianran Zhang et al. “Understanding electrode materials of rechargeable lithium batteries via DFT calculations”. In: *Progress in Natural Science: Materials International* 23.3 (2013), pp. 256–272.

[36] Materials Project. *Materials Project LiFePO4*. [Online; accessed 26-February-2020]. 2020. URL: %5Curl%7Bhttps://materialsproject.org/materials/mp-585319%7D.

[37] Kang Xu. “Nonaqueous liquid electrolytes for lithium-based rechargeable batteries”. In: *Chemical reviews* 104.10 (2004), pp. 4303–4418.

[38] Hansu Kim et al. “Metallic anodes for next generation secondary batteries”. In: *Chemical Society Reviews* 42.23 (2013), pp. 9011–9034.

[39] Yayuan Liu et al. “Lithium-coated polymeric matrix as a minimum volume-change and dendrite-free lithium metal anode”. In: *Nature communications* 7 (2016), p. 10992.

[40] YJ Zhang et al. “An ex-situ nitridation route to synthesize Li3N-modified Li anodes for lithium secondary batteries”. In: *Journal of Power Sources* 277 (2015), pp. 304–311.

[41] Lu Li et al. “Self-heating–induced healing of lithium dendrites”. In: *Science* 359.6383 (2018), pp. 1513–1516.

[42] Hongkyung Lee et al. “Suppressing lithium dendrite growth by metallic coating on a separator”. In: *Advanced Functional Materials* 27.45 (2017), p. 1704391.

[43] Thomas B Reddy. *Linden’s handbook of batteries*. Vol. 4. Mcgraw-hill New York, 2011.

[44] Hyun Deog Yoo et al. “Mg rechargeable batteries: an on-going challenge”. In: *Energy & Environmental Science* 6.8 (2013), pp. 2265–2279.

[45] Qingfeng Li and Niels J Bjerrum. “Aluminum as anode for energy storage and conversion: a review”. In: *Journal of power sources* 110.1 (2002), pp. 1–10.

[46] Richard Van Noorden. “The rechargeable revolution: A better battery”. In: *Nature News* 507.7490 (2014), p. 26.

[47] Matylda N Guzik, Rana Mohtadi, and Sabrina Sartori. “Lightweight complex metal hydrides for Li-, Na-, and Mg-based batteries”. In: *Journal of Materials Research* 34.6 (2019), pp. 877–904.

[48] John Muldoon et al. “Electrolyte roadblocks to a magnesium rechargeable battery”. In: *Energy & Environmental Science* 5.3 (2012), pp. 5941–5950.

[49] Doron Aurbach et al. “Nonaqueous magnesium electrochemistry and its application in secondary batteries”. In: *The Chemical Record* 3.1 (2003), pp. 61–73.

[50] Doron Aurbach et al. “A comparison between the electrochemical behavior of reversible magnesium and lithium electrodes”. In: *Journal of power sources* 97 (2001), pp. 269–273.

[51] JS Gnanaraj et al. “Improving the high-temperature performance of LiMn2O4 spinel electrodes by coating the active mass with MgO via a sonochemical method”. In: *Electrochemistry Communications* 5.11 (2003), pp. 940–945.

[52] Ran Attias et al. “Anode-electrolyte interfaces in secondary magnesium batteries”. In: *Joule* 3.1 (2019), pp. 27–52.

[53] Yanguang Li and Jun Lu. “Metal–air batteries: will they be the future electrochemical energy storage device of choice?” In: *ACS Energy Letters* 2.6 (2017), pp. 1370–1377.

[54] MK Aydinol and G Ceder. “First-Principles Prediction of Insertion Potentials in Li-Mn Oxides for Secondary Li Batteries”. In: *Journal of the Electrochemical Society* 144.11 (1997), p. 3832.

[55] Venkat Srinivasan. “Batteries for vehicular applications”. In: *AIP conference proceedings*. Vol. 1044. 1. American Institute of Physics. 2008, pp. 283–296.

[56] John P Perdew and Mel Levy. “Physical content of the exact Kohn-Shamw orbital energies: band gaps and derivative discontinuities”. In: *Physical Review Letters* 51.20 (1983), p. 1884.

[57] John P Perdew. “Density functional theory and the band gap problem”. In: *International Journal of Quantum Chemistry* 28.S19 (1985), pp. 497–523.

[58] Anubhav Jain et al. “A high-throughput infrastructure for density functional theory calculations”. In: *Computational Materials Science* 50.8 (2011), pp. 2295–2310.

[59] Maarten De Jong et al. “Charting the complete elastic properties of inorganic crystalline compounds”. In: *Scientific data* 2.1 (2015), pp. 1–13.

[60] Tom M Mitchell. *Machine learning*. 1997.

[61] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[62] Fangyu Cai. *MNIST Reborn, Restored and Expanded: Additional 50K Training Samples*. [Online; accessed 22-March-2020]. 2019. URL: %5Curl%7Bhttps://miro.medium.com/max/1060/1*VAjYygFUinnygIx9eVCrQQ.png%7D.

[63] James E Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of computational statistics: concepts and methods*. Springer Science & Business Media, 2012.

[64] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2014.

[65] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[66] Manuel Fernández-Delgado et al. “Do we need hundreds of classifiers to solve real world classification problems?” In: *The journal of machine learning research* 15.1 (2014), pp. 3133–3181.

[67] Ioannis Tsamardinos et al. “An Automated Machine Learning Architecture for the Accelerated Prediction of Metal-Organic Frameworks Performance in Energy and Environmental Applications”. In: *Microporous and Mesoporous Materials* (2020), p. 110160.

[68] M Attarian Shandiz and R Gauvin. “Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries”. In: *Computational Materials Science* 117 (2016), pp. 270–278.

[69] Christian Robert. *Machine learning, a probabilistic perspective*. 2014.

[70] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[71] *SklearnCross-val*. `https://scikit-learn.org/stable/modules/cross_validation.html`. Accessed: 2020-03-23.

[72] HE ELECTRO and E MIC. “Materials Scientists Look to a Data-Intensive Future”. In: ().

[73] Radislav Potyrailo et al. “Combinatorial and high-throughput screening of materials libraries: review of state of the art”. In: *ACS combinatorial science* 13.6 (2011), pp. 579–633.

[74] Jana Kalawoun et al. “From a novel classification of the battery state of charge estimators toward a conception of an ideal one”. In: *Journal of Power Sources* 279 (2015), pp. 694–706.

[75] Ephrem Chemali et al. “State-of-charge estimation of Li-ion batteries using deep neural networks: A machine learning approach”. In: *Journal of Power Sources* 400 (2018), pp. 242–255.

[76] Xiaosong Hu et al. “Battery health prognosis for electric vehicles using sample entropy and sparse Bayesian predictive modeling”. In: *IEEE Transactions on Industrial Electronics* 63.4 (2015), pp. 2645–2656.

[77] Stefano Ermon et al. “Learning policies for battery usage optimization in electric vehicles”. In: *Machine learning* 92.1 (2013), pp. 177–194.

[78] Fei Zhou et al. “First-principles prediction of redox potentials in transition-metal compounds with LDA+U”. In: *Physical Review B* 70 (Dec. 2004), p. 235121. ISSN: 1098-0121. DOI: 10.1103/PhysRevB.70.235121. URL: http://link.aps.org/doi/10.1103/PhysRevB.70.235121.

[79] Stefan Adams and R. Prasada Rao. “High power lithium ion battery materials by computational design”. In: *Physica Status Solidi (a)* 208.8 (Aug. 2011), pp. 1746–1753. ISSN: 18626300. DOI: 10.1002/pssa.201001116. URL: http://doi.wiley.com/10.1002/pssa.201001116.

[80] Austin D Sendek et al. “Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials”. In: *Energy & Environmental Science* 10.1 (2017), pp. 306–320.

[81] Rajendra P Joshi et al. “Machine Learning the Voltage of Electrode Materials in Metal-ion Batteries”. In: *ACS applied materials & interfaces* 11.20 (2019), pp. 18494–18503.

[82] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[83] John W Backus and William P Heising. “Fortran”. In: *IEEE Transactions on Electronic Computers* 4 (1964), pp. 382–385.

[84] Travis E Oliphant. *A guide to NumPy*. Vol. 1. Trelgol Publishing USA, 2006.

[85] Felipe Pezoa et al. “Foundations of JSON schema”. In: *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2016, pp. 263–273.

[86] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 51–56.

[87] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[88] Maarten de Jong et al. “Charting the complete elastic properties of inorganic crystalline compounds”. In: *Scientific Data* 2 (Mar. 2015). DOI: 10.1038/sdata.2015.9. URL: http://perssongroup.lbl.gov/papers/sdata2015-elasticprops.pdf.

[89] Shyue Ping Ong et al. “The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles”. In: *Computational Materials Science* 97 (2015), pp. 209–215. DOI: 10.1016/j.commatsci.2014.10.037. URL: http://dx.doi.org/10.1016/j.commatsci.2014.10.037.

[90] Daniele Ongari et al. “Accurate characterization of the pore volume in microporous crystalline materials”. In: *Langmuir* 33.51 (2017), pp. 14529–14538.

[91] Anthony K Rappé et al. “UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations”. In: *Journal of the American chemical society* 114.25 (1992), pp. 10024–10035.

[92] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[93] Pierre Geurts, Damien Ernst, and Louis Wehenkel. “Extremely randomized trees”. In: *Machine learning* 63.1 (2006), pp. 3–42.

List of Figures

List of Tables