UiO : Department of Physics
University of Oslo

Sondres master utkast 10.03.2020

Sondre Torp - Sondrt@student.matnat.uio.no

Abstract

This is the Abstract!

Acknowledgements

I like to acknowledge ...

Sabrina Sartori

Tor Svendsen Bjørheim

UOC

prof. George E. Froudakis, prof. George Fanourgakis

m.m.

titles?

Contents

List of Figures

List of Tables

Part I

Introduction

1 Introduction and Overview

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

In the introduction a general background for this work will be given. An overview to earlier work in the field of ML on batteries and other materials, will be given, which will include state-of-the-art relevant for this master thesis. The chapter includes a motivation, the scope of the thesis and its structure.

1.1 Earlier work

The field of material science in blooming in an immaculate rate due to computing power being cheaper and more available then ever before. When the Human Genome Project started in the 1990 the main problem was lack of computational power, but due to Moore's law the sequencing of all 3 billion letters of DNA is not even a challenge in todays standards. While in the field of material science, researchers have know how to, in theory, simulate materials since the early 1900, due to the discovery of quantum mechanics. The problem being that a advanced material have an order of 10^{23} electrons that demands computational methods to simplify the problem [1]. And the gap between the energy storage needed and what state-of-the-art systems are providing is increasing [2].

There has been a constant drive for battery innovation in the last 40 years. For example, 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 [3][4], which researchers are working on [5] [6] [7] [8]. In recent years a desire to move forward towards an ultimate energy density technology has forced researchers to evaluate technologies beyond Li-ion batteries, other metals such as magnesium and aluminum are pointed out [2] [9]. Aluminum and magnesium are considered, among other things, due to their abundance. In the case of aluminum it has a high theoretical voltage, specific energy, and is the most abundant metal in the world, but it is hindered by a oxide layer on its surface [10], but solutions to this problem is being offered for large-scale applications [11]. Magnesium has attracted increased attention due to its higher volumetric capacity than lithium (i.e. $3832\mathrm{mAhcm}^{-3}$ vs $2061\mathrm{mAhcm}^{-3}$) and being the fifth most abundant element [12]. Mg is not competitive with Li metal on both specific capacity and redox potential ($700\mathrm{mV}$ less), but dendrite formation is absent, which alleviates the safety concerns [13]. Even still there are several roadblocks ahead when looking at the possible electrolytes. One of the bigger once

are the corrosive property of magnesium organohaloaluminates. Another is the unique electrochemistry which prohibits its reversible deposition in aprotic solvents containing currently commercial ionic salts such as magnesium bisimide or magnesium perchlorate. Reduction of these electrolytes creates blocking surface layer that inhibits deposition and conduction of magnesium ions [14] [15].

In the field of computational material design a subfield called 'high-throughput' (HT) computational material science is on the rise [16]. This burgeoning 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 analyse the data and find materials with desired properties. This method should continuously be validated by comparing the calculated, or predicted, values on real (already known) materials, and later also on new/hypothetical materials, to create a feedback loop to further improve the algorithm [17]. A computational HT consists of three tightly connected steps: Virtual material growth, rational material storage, material characterizations and selection. This work is based in the second and third step, rational material storage and material characterization on intercalation type batteries.

Several studies on batteries have applied machine learning (ML) and different degrees of HT, in particular on electrical vehicles and how to estimate their state of charge accurately and in the direction of energy system management [18] [19] [20] [21]. In the direction of predicting battery properties, only a handfull of studies were found. Shandiz and colleagues [22] used classification methods to determine the crystal system of silicate cathodes. They found that the classifier random forrest gave the highest accuracy of prediction, and that there is a strong correlation between the three major crystal system (monoclinic, orthorhombic and triclinc), that they did their predictions on, and other features of cathodes. They used the online database Materials project [23] [24] [25]. Sandek and colleagues [26] constructed a classification model using logistic regression to find possible solid state electrolytes for lithium ion batteries. Finding that simple atomistic descriptors alone, did not function to get useful prediction. They found that if they combined these with structural and electronic information, in the Materials project database,

Other areas where ML have shown promise are in the field of Nanoporous Materials [27]. Where a set of new descriptors for predictions on methane adsorption was proposed. Fanourgakis and colleagues combined structural features, such as helium void fraction, surface area, and pore volume, with other descriptors found by using probe atoms of various size on MOFS they could predict the methane uptake capability even under low pressure. Which later lead to a more general application of ML on nanoporous materials [28]. Where 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. This also showes that these predictors might be applicable on other type of materials, due to the different nature of MOFs and COFS.

1.2 Scope of the thesis

Batteries are vastly complex and much efforts have been devoted to their development. Yet, with all the efforts put in to electrochemical ask Sab what she wrote cells, there are still a never ending chase for batteries that can push the limits of their properties even further. The demand for better batteries are growing faster then ever. The global electric car fleet, for instance, exceeding 5.1 milion[29], almost doubling the number of new electric car registrations in the last year. And if we follow the policies of the EV30@30 Scenario which aims to reach a 30% market share for EVs in all models except two wheelers by 2030, due to one quarter of global greenhouse gas(GHG) emissions comes from this sector. The EV sales per year are then predicted to be more then 43 milion, and the stock numbering more than 250 milion.

This work develops 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 base properties taken from existing databases, we propose a set of predictors to predict the properties of new, undiscovered electrodes, or even new properties in already well known electrodes.

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. In this work we will focus on the base properties; voltage, energy density, specific energy, and the physical stability of the materials.

The idea behind the design of a battery is fairly simple, due to their similar procedure to create electricity. Two of the most important parts are the Anode and the Cathode. A cell produces electricity when placed into a medium that has conductive properties, namely the electrolyte.

Today some of the main methods for theoretical advances in battery science are density functional theory (DFT), molecular dynamic simulations and elaborate

1.3 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. Due to the complexity of the chemical processes involved, it is of high importance to be able to develop predictive modeling methods to improve the search for better compositions and performances. elaborate more on predictive modeling ...

1.3.1 Research Question

How to improve batteries?

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

RQ2: Which ML method would be the most optimal for such a search?

RQ3: What predictors are the most suited for such a task, and which would yield the most efficient training.

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

1.3.2 Approach

To answer the RQs we need to establish a database that is comprehensive enough to be of value for the machine learning method we decide to use and has the relevant information needed to get good predictions.

The question of what ML method that would be the most optimal one is a difficult question, so we approached it with the question; what machine learning method is more likely to yield correlation in our data?

The choice of features examined in this work is inspired by an extensive survey done on a similar project especially in the field of Metal Organic Framework(MOFS) preformed by collaborators from the university of Crete [27] [28], dictated, to some degree by the lack of more data, and by the data we had.

In order to evaluate the effect of different features, a prediction approach using principle component analysis was chosen. First we used physical descriptors, such as geometrical properties of the unit cell. This because it was greatly efficient in similar studies on MOFS, and it is straightforward. Other descriptors are needed, and the void fraction seemed like the next obvious one, it is both geometrical and easy to obtain. We still needed a more chemical approach without doing big DFT type calculations, so we tried to make a atomic property weighted radial distribution function, to include tablature values like electronegativity, van der waals volume and magnetization. Before lastly using the already known results, or targets, to make predictions on the other targets.

All of these methods were tested alone, and up against each other both for a smaller database of Mg-ion intercalation type batteries, and a bigger database of Li-ion intercalation type batteries. And the results of these two different databases might not be purely comparable, but due to the algorithm mainly using fundamental properties we will compare the results and look for trends as our database increases in size.

1.4 Structure of the thesis

First the most essential concepts from the fields of batteries, machine learning, and work already explored on these two fields conjoined, are introduced. Then the method will be ex-

plained before presenting and discussing our results so far, before trying to put this all into perspective and conclude. All the work done is available in github.

Part II

Foundations

2 Batteries

In this section, a brife summery of some of the vocal points in the history and evolution of batteries, with a following consideration of lithium ion batteries and magnesium batteries, and their roll in todays battery market. The basic principles of batteries will be explained, with a special focus on electrodes which are the part of the battery that this work will be centered. Some of the more essential properties related to the chemistry of this work will also 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 [17]. 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 by Alessandro Volta (1745 - 1827). Who built the first known battery in the year 1800 [30]. His invention consisted in the voltaic, pile with zinc and copper plates stacked on top of each other and sheets of bine-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. See figure: 1.

It took almost 40 years before the British inventor John Frederic Daniell Voltas continuation with the discovery of the Daniell cell [32] in 1836. The Daniell cell 2.1, as sketched below 2, is constructed with two half cells, one with a zinc electrode in a zinc sulfate dissolution, and another copper electrode in a copper sulfate solution. These half cells are connected by a salt bridge. This cell could give a voltage of $1.1\,\mathrm{V}$ through the reaction shown below.

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

In 1859 the French physicist Gaston Planté constructed the first led-acid battery. 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 stripps, so that the plates would not touch. The lead-acid battery was special due to the electrolyte being a active part of the chemical reaction in the battery. The electrodes in the battery 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 discharged. The charge

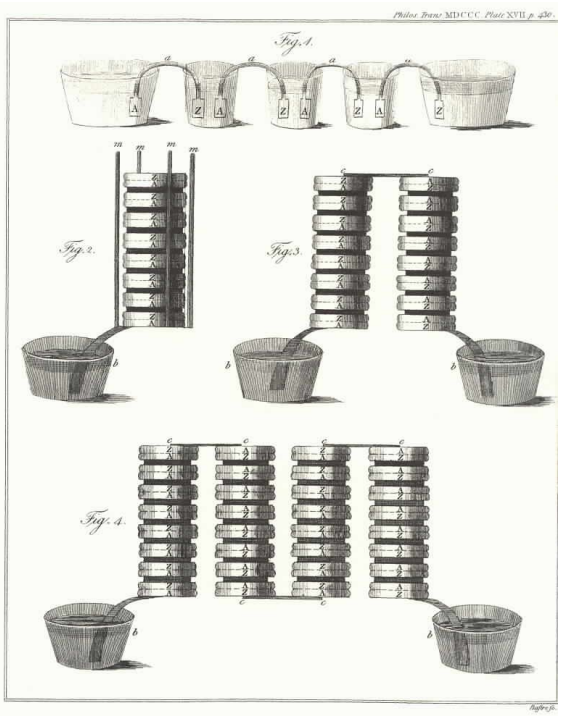


Figure 1: A voltaic pile, the first battery [31]

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.

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

The open circuit voltage (V_{OC}) for a led-acid battery are approximately $2\,\mathrm{V}$. It is custom to attach these batteries in series to attain a higher voltage, typical 6 or $12\,\mathrm{V}$. These batteries have a shelf- and cycle life of more than 10 years or 1000-2000 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 its weight, they are also renowned for their high environmental impact. Therefor, one of the many goals of battery producers is to replace led-acid batteries with higher preforming alternatives.

Nickel-cadmium (NiCd) batteries were first described by the swede Waldemar Jungner in 1899 [34]. These batteries rose fast 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 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 several similarities with the NiCd batteries. The main difference is the anode with Cd, while

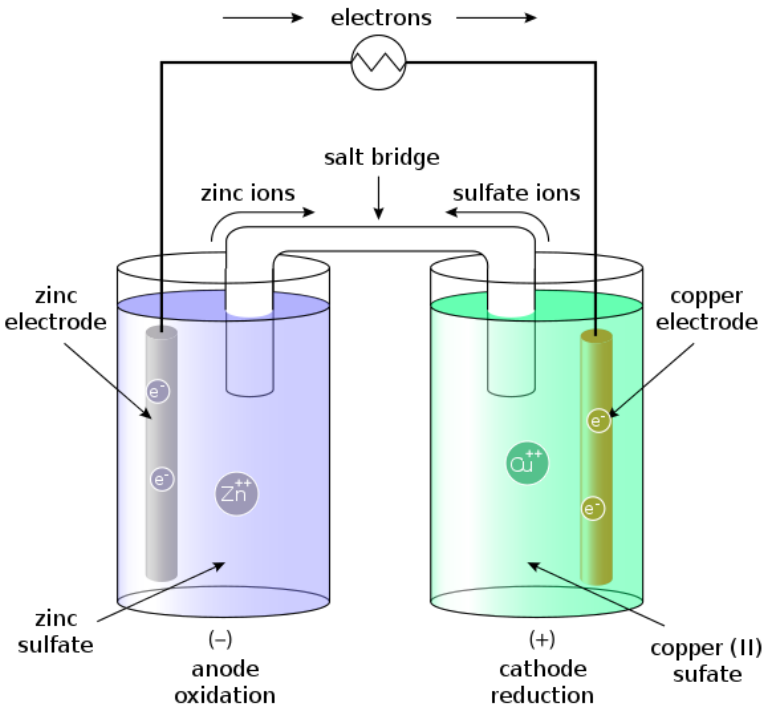


Figure 2: A draft of a Daniell cell. The anode is a piece of zinc and the cathode a piece of copper. The salt bridge transports ions between the solutions and the electrons moves through an external circuit [33].

in NiMH it 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.

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

A primary cell is a non-rechargeable battery. These batteries are normally 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 a electrolyte of potassium hydroxide. A typical alkaline battery delivers a nominal cell voltage of $1.5\,\mathrm{V}$. The overall reaction for a alkaline battery is shown below (2.5).

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

2.2 Lithium based batteries

The intercalation electrodes for lithium and other alkaline metals were discovered in 1975 by Micheal Stanley Whittingham [35]. This led to the first lithium batteries with titanium disulfide (TiS_2) as the cathode and metallic lithium as the anode. TiS_2 - structure is divided into layers and lithium-ions are inserted or extracted from interstitial space between atomic layers of the active materials, 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). While applying an over-potential to the material, or charging, the lithium-ions move out of the TiS_2-structure and titanium oxidzes 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.

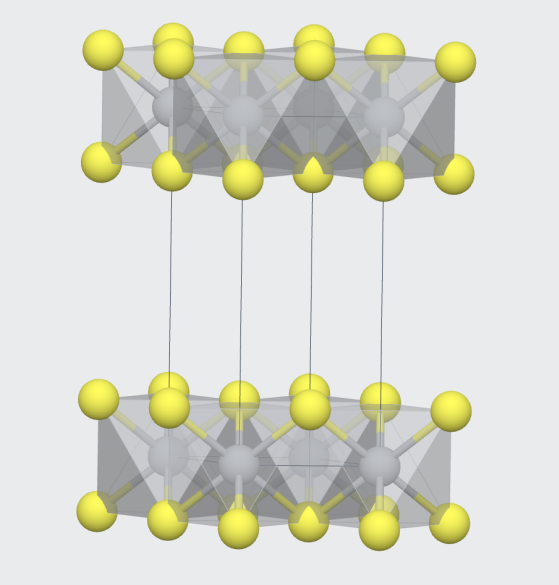


Figure 3: The two-dimensional trigonal omega-like structure of TiS_2. Observing from a slight angle along the b-axis. The titanium in grey, are occupying the octahedral holes of sulfur, in yellow. Lithium-ions would intercalate into the space between the TiS_2 layers [36].

Such layered structures allow their reversible behavior. In 1980 John B. Goodenough introduced $LiCoO_2$ (LCO) as the cathode material for lithium batteries, and earned him, M. Stanley Whittingham and Akira Yoshino the Nobel Prize in Chemistry in 2019 [37]. Goodenough and colleagues found a current density of up to $4\,\mathrm{mAcm^{-2}}$ [38] [39]. Even though the properties where exceptionally good at the time, the batteries were still not commercialized due to metallic lithium being to 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 provides good electrical performance, are relatively safe, easy to prepare, and is not especially sensitive to process variation and moisture. The metallic lithium anode was sub-

stituted with 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 has a anode of metallic lithium. This nomenclature is transferable to other type of batteries like, magnesium-ion/magnesium batteries.

Figure 4 shows a typical of a lithium-ion battery with $LiCoO_2$ as the cathode and graphite as the anode. During discharged 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 separat external circuit, where the electrical energy can be extracted. When charged a overpotential is applied the reaction is reversed. The overall reaction is shown in equation: (2.6)

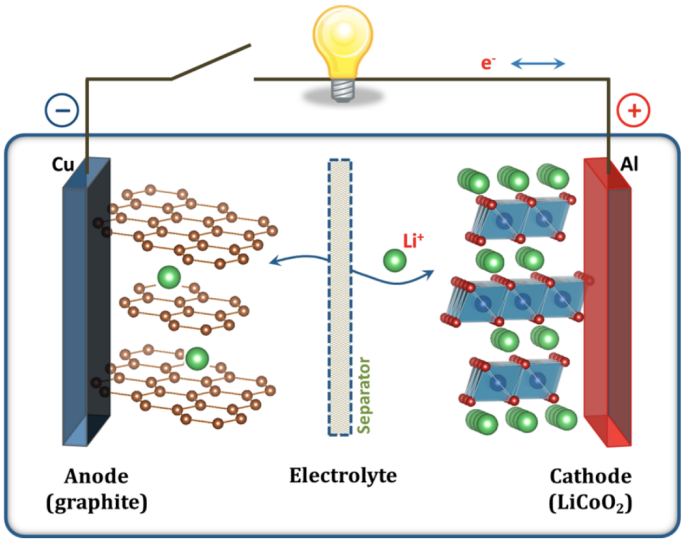
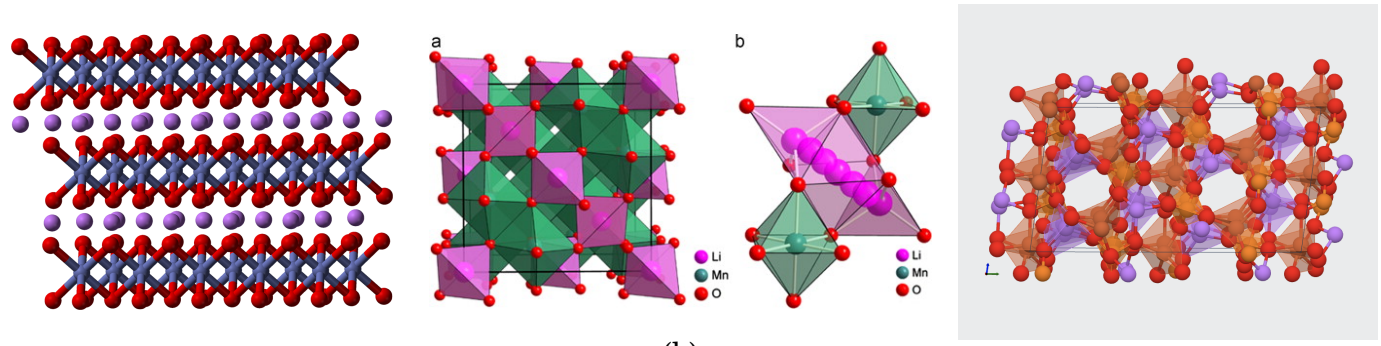


Figure 4: Schematic illustration of the first Li-ion battery $LiCoO_2/Li^+$ electrolyte/graphite [40].

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

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 $LiFePO_4$ (**LFP**). $LiMn_2O_4$ is a good ionic conductor due to the structure having channels in all three dimensions where lithium can be transported 5b. $LiFePO_4$ has a lower ionic conductivity of the two, $LiMn_2O_4$, due to only having channels in one dimension, see 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 a) $LiCoO_2$ [41] b) $LiMn_2O_4$ [42], c) $LiFePO_4$ [43]. The layers in $a)$, the 3-dimensional canals in $b)$ and the 2-dimensional channles in $c)$, 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. To improve the cathode material is therefore high in priority among many research groups. Other recent anode materials are $Li_{4/3}Ti_{5/4}O_4$ spinel, which has a lower specific capacity and capacity density than graphite, but has a longer cycle life and good termal stability characteristics. Nanostructured $Sn-Co-C$ alloys commercialized in 2005 by Sony and Si-based negative electrodes seems 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; they have sealed cells - which means that their is no requirement of maintenance, long shelf and cycle life, low self discharge rate, high energy efficiency, high energy density, high-rate and high-power discharge capability, no memory effect, many possible chemistries offer design flexibility. While the drawbacks of some Li-ion batteries are; moderate initial cost, degeneration when discharged below $2\,\mathrm{V}$, degrades at high temperature (above $65°\mathrm{C}$ they can permanently lose capacity), their need for protective circuitry, capacity loss and potential for thermal runaway when overcharged and when crushed. They can also become unsafe if rapidly charged at sub zero temperatures.

There is 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 [44]. 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 them showing high ionic conductivity, and in some cases electrochemical properties that enable battery reversibility.

2.3 Magnesium based batteries

The abundance of magnesium in the earth's crust, combined with its low atomic weight, low cost, and electrochemically active nature, makes magnesium a good candidate for battery applications. It can serve as a potential negative electrode with its electrochemical potential of $-2.37\,\mathrm{V}$, and it is environmentally friendly.

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 [2].

There is an ongoing search for high performance cathode magnesium materials for the realization of a practical, rechargeable Mg battery. The Mg^{2+} showes 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 [11]. 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 in the research stage [45].

2.4 Cell operation principles and design

Batteries and fuel cells 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 gain or lose an electron by changing the oxidation number.

A 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 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 normally 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. The electrolyte is typically a liquid, to impart ionic conductivity. It can be a solid, but this is less common. The cell will produce electricity when the circuit is complete. The electrolyte can, in some designes, act as both the electrolyte and the 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. [46]

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 with the application of an external electrical power source, so that it creates an over-potential - 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 anode and cathode are changed, the cell will yield more or less 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.

2.4.1 The manufacturing process of an Alkaline battery

In alkaline batteries the cathode is used both as the container and the cathode. Manganese dioxide, graphite, and the electrolyte are mixed and ganulated and pressed into hollow cylinders called preforms. The preforms can be stacked or replaced by an extruded ring of the same material. The preforms are then combined with nickel-plated steal. These are then the containers of the battery.

A separator is then soaked in the electrolyte solution and inserted between the cathode and where the anode is supposed to be. The anode is then put into the can, it is a gel composed of zinc powder and other materials like potassium hydroxide electrolyte. The gel does not fill the whole can, so that their is room for the chemical reaction to occur.

The can is then sealed with three connected components. The first being a current collector, going to thirds of the way trough the anode with a plastic seal at the top before it is all closed by a metal cover with a metal cap. The current collector goes all the way through the plastic seal and is connected to the metal cap. The seal is thinner in some places, in case of gas build ups. In some battery designes, wax are used instead of a plastic seal, so excessive gas can push through the wall. Lastly a label is attached to the battery, with the necessary information about the battery.

The battery is then taken through a complex quality control, to certify the batteries ability to resist corrosion, maintain a good shelf life, usage life, and other factors.

useless section?

2.4.2 Theoretical Cell voltage, Capacity, specific Energy, Energy density

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:

$$aA + ne \leftrightharpoons cC$$

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

$$bB \leftrightharpoons dD + ne$$

with the overall reaction being:

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

Whenever there is a reaction, there is a decrease in the free energy of the system, namely 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. 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 is determined by the type of active material in the cell, and can be calculated from the free energy. The standard potential of a cell can be calculated form the standard electrode potential.

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

Direct measurements of the absolute electrode potential is very hard, so a reference point, the standard potential of the H_2/H^+ is taken as zero and all other standard potentials are referred to this potential.

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

$$E = E^0 - \frac{RT}{nF} \ln \frac{a_C^c a_D^d}{a_A^a a_B^b} \tag{2.8}$$

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 two chemical reactions, redox reaction, in the electrodes that creates a potential gap between the electrode and the electrolyte, when a outer circuit is connected this gap is lowered, but due to the reaction rates going up, the potential gap is maintained.

Electrode processes To characterize an electrode it is custom to use both chemical and electrical changes. The electrode reaction may be simple, but despite this simplicity the overall process involves many more steps, which makes it more complex. rewirte. Transport of electroactive species prior to the electron transfer step. Adsorption of the electroactive material may be involved. Chemical reaction is often involved in the overall electrode reaction. Any reaction rate is dictated by its slowest step. what are you trying to say.

Experimentally, it is shown that there is an exponential relationship between current and applied voltage

The theoretical *capacity* is determined by the amount of active material in the cell. If the calculation 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 $0.394\mathrm{Ah/g}$.

$$\mathrm{Ze} + \mathrm{Cl_2} \rightarrow \mathrm{ZnCl_2}$$
$$1.22\mathrm{g/Ah} + 1.32\mathrm{g/Ah} = 2.54\mathrm{g/Ah}$$

The active materials of the electrodes allow the reversible uptake and release of Mg, or Li ions. This may happen by; movement of the ions; into, i.e. *intersection* or *intercalation* or out of, i.e. *extraction* or *deintercalation*, their chemical structures, *phases*. This is done by conversion of the materials between ion poor and rich i.e. *alloying*, or rich and poor, e.g. *dealloying* phases, this can also be done by conversion of the electrode material into other more ion rich/poor chemical forms or mixtures. Referred to as *conversion* or *displacement* reaction, with the average ion content of the entire electrode varying.

The total Li or Mg content in the electrodes will thus 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, due to the *database*, more on this later.

Specific Energy

Specific Energy, or gravimetric energy density, defines battery capacity in weight, energy density, or volumetric energy density, defined as:

$$\text{Watthours/gram} = \text{Voltage} \times \text{Ampere-hours/gram} \tag{2.9}$$

theoretical specific energy - based on the active anode and cathode only theoretical specific energy of a practival battery, accounting for electrolye and nonreactive components The actual specific energy of these battereis when discharged at 20C under optimal discharge conditions.

2.5 Battery properties

Batteries are characterized according to several different properties, both chemical and electrical. Some of the most important once are the physical properties energy, energy density, capacity, power, and current (I). These relate to each other as shown in equation 2.10, 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

$$V \cdot C = E \tag{2.10}$$

$$\frac{E}{Volume} = U \tag{2.11}$$

$$V \cdot I = P \tag{2.12}$$

$$W \cdot t = E \tag{2.13}$$

Average voltage The voltage of a battery is determined by the types of active materials that are used. The cell voltage are also limited by concentration and temperatures, as expressed by the Nernst equation.

Capacity can be defined as:

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

And is the i number of electrons or cations exchanged between the negative and positive electrodes, i.e. how much charged 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. The capacity is normally expressed as capacity per mass (specific or gracimetric capacity), $\mathrm{Ah/kg}$, or per volume (volumetric capacity), due to its dependency on the amount of active material. Theoretically, capacity is 1 gram equivalent weights of the active material (in grams) divided by the number of electrons in the reaction.

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

The battery can deliver power which is defined as:

$$P(t) = V(t)I(t) \tag{2.15}$$

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

$$W = \int P(t) \cdot dt = \int V(t)I(t) \ \cdot dt \tag{2.16}$$

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

Energy density

Gravimetric 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.

2.6 Cell limitations & definitions

In this paper, especially under the section on general properties of battery, an amount of technical terms related to material details, from is used. Here I will clarify what we mean by these terms.

2.6.1 Polarization

Polarizability is a tabulated atomic properties, it is the ability to form instantaneous dipoles(REF), and is defined as:

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

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.

2.7 Battery chemistries

2.8 Intercalation batteries and why Li

The current Li-ion batteries offer some of the best combination of high specific energy, energy density, long cycle life and high-power capability, among the rechargeable battery technologies. They have dominate the worldwide market in portable and consumer electronic equipment, electric vehicles, space applications, as well as electrical energy storage. Li-ion batteries have minimal side reactions when a Li ion intercalates into the cathode/anode materials. They exhibit limited self-discharge, and no memory effects that limit energy density after many cycles. Their energy efficiency may be further enhanced by lowering the internal resistance of the battery, and Li-ion batteries as a result, receive considerable attention at both fundamental and applied research levels.

The development of stable novel materials is the key to the successful development of novel and advanced rechargeable batteries. Current research and development has focused on upgrading the energy density of Li-ion batteries.

Most practical rechargeable batteries deliver capacities and energy densities far below their theoretical values(med kilder)due to limited utilization efficiency of the active materials that participate in electrochemical reactions. The major reasons for such limitations include effects that result from slow electrode process kinetics with high polarization and low ionic diffusion or electronic conductivity rates, particularly at the electrolyte–electrode interfaces. Material stability issues caused by a low Li content can also impact on its degree of charging. Therefore, the improvement in existing rechargeable battery systems involves exploring key materials and

focusing our attention on the atomic, ionic, or molecular diffusion and transport. Charge transfer, the optimization of surface and interface structure, and the regulation of electrochemical reactions within Li-ion systems may pave the way for improved (i) capacity, and energy and power density, (ii) reactivity, reversibility, and structural stability during charge– discharge cycles, (iii) ionic diffusion and electronic transfer at high charge–discharge rate, and (iv) lower cost, increased safety and environmental compatibility

inkluder litt om dette? og med bilder kanskje?

The schematic diagram of the current Li-ion battery based on a carbon based anode (LixC6), cathode (LiCoO2), liquid electrolyte (LiPF6 dissolved in a mixture of EC and DMC or equivalent), and separator is shown in Fig. 2.9 In the foreseeable future, Li-ion batteries will be the most practical solution to a wide range of electrical energy storage applications.10

. The majority new cathode materials for Li-ion batteries under research and development are transition metal oxides, which tend to provide lower discharge potential as the electric-capacity density increases. Carbon-based materials (usually graphite) are currently used as anode materials in Li-ion batteries. The other variety of carbon-based materials and pure Li metal are currently proposed as alternate anode materials, but many need further improvement with respect to electrode potential and charge–discharge cycle life concerns

"[47]

Working principle of Li-ion cell.

When a Li-ion battery discharges, a Li^+ moves from the anode (i.e. graphite) to the cathode (i.e. $\mathrm{LiMO_2}$ where M is a transition metal), through the electrolyte that commonly is Li^+ - containing salt. The reaction, as discussed above 2.7, are:

$$\mathrm{LiMO_2} \leftrightharpoons \mathrm{Li_{1-x}MO_2} + \mathrm{xLi^+} + \mathrm{xe^-} \text{ (cathode)}$$
$$\mathrm{xLi^+} + \mathrm{xe^-} + \mathrm{xC_6} \leftrightharpoons \mathrm{xLiC_6} \text{ (anode)}$$

The overall reaction:

$$\mathrm{LiMO_2} + 6\,\mathrm{C} \leftrightharpoons \mathrm{Li_{1-x}MO_2} + \mathrm{xLiC_6}$$

The anode is graphite, thus there is noe metallic Li, which makes the Li-ion battery less reactive, therefor safer, and the graphite anode offers a longer cycle life than their Li-metal counterpart. To progress the performance of Li-ion batteries a few design changes are needed; A cathode with a chemical potential that matches the electrolytes highest occupied molecular orbital, and an anode matching the lowest unoccupied molecular orbital of the electrolyte. A non-aqueous electrolyte of high Li^+ ion conductivity under practical temperatures.

2.9 Electrodes and features

In this section the features used in ML as predictors will be introduces. First will the pair properties be introduced, before going into the more electrode specific features.

As a general note. These features 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. Nice to give this note?

Average Voltage

The theoretical voltage and capacity of a cell are function of the anode and cathode materials, with the composition of the electrolyte, and at the temperature of $25°\mathrm{C}$. write better

The active materials contained in the cell determines the standard potential, E^0, which can be calculated from the free-energy. The standard potential of a cell can be calculated from the standard electrode potential:

$$\text{Anode(oxidation potential)} + \text{cathode (reduction potential)} = \text{standard cell potential} \quad (2.17)$$

The cell voltage is also dependent on other factors including concentration and temperature, as expressed by the nernst equation. (REF) *Average Voltage* as we use, is defined as the voltage average during the discharge. It is lower then the theoretical voltage.explain why this is, ref

Physical stability

What we refer to as Physical stability is Energy above hull. 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.

2.9.1 Properties of materials

Are These (or some of them) theoretically computed properties. If yes provide some details about how they computed (i.e. DFT methods)

Total *Magnetic Moment* (μ_B) is calculated for the unit cell within the magnetic ordering provided.

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

Energy Above Hull per Atom The energy of decomposition of the material into the set of most stable materials at this chemical composition, in eV per atom. 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.

Density, here defined as the calculated bulk crystalline density, typically underestimated due to the calculated cell volume being overestimated on average by $3\%(+-6\%)$.

The *Band Gap*s are calculated a little different. In general, band gaps are computed with common exchange-correlation functionals such as the LDA [48] and GGA are severely underestimated [49]. Typically the disagreement is reported to be $\tilde{5}0\%$ in the literature. Some internal testing by the Material Project supports these statements; typically, they found that band gaps are underestimated by $\tilde{4}0\%$. We additionally find that several known insulators are predicted to be metallic.

Cycle life

Rate capability

RC

Self discharge

SD

Energy per atom

EpA

volume

Volume of the unit cell defined as

Formation energy per atom

Fepa

Band gap

The band gaps of a solid is simply the range of energies an electrode in a solid can not have. While the bandstructurs. How much to include? Should I here have a page on quantum physics and the bandstructure? - Only include relevant stuff.

Total magnetization

Tm

Elasticity

E

Porous Electrodes

In a battery, the reactant is supplied from the electrolyte phase toe the catalytic electrode surface. Electrodes are often composites made of active reactants, binders and fillers, in batteries. To minimize the energy loss of both activation and concentration polarizations at the electrode surface and to increase the electrode efficiency or utilization, it is often preferred to have a large electrode surface area. This can be done by have a porous electrode design. A porous design can provide an interfacial area per unit volume that is considerable higher den that of a planar electrode.

A porous electrode is a electrode that concists of porous matrice 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.

2.10 Mg- and Li- batteries: State-of-the-art

3 Machine Learning

In this chapter we summarize some concepts of machine learning and related ideas. The first section introduces the basic ideas behind machine learning and *some of the best known examples* will be presented. Secondly the concepts of supervised and unsupervised learning will be presented with a clarification on the difference between regression and classification problems, so that we can define 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, mean square error(MSE), root mean square error (RMSE) and R-squared(R^2).

Before we round of this section with a brief explanation on the role of data, how features can affect the effectiveness of a model, and finalizing with the concepts of over- and under-fitting, 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..."([50]). 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 Example time!

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". (PICTURE TIME!)

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 one human being. Therefore a model which can recognize these digits would be useful. A model that takes a picture of a digit and outputs that digit in a way that is recognizable for a machine, that is, 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([51]). 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}

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 is nothing more than a grid of numbers that represent how dark a pixel is. So each picture contains 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 A. A 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 prosess 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. [52]

In the case of supervised learning one knows the answer to a problem, and let the computer deduce its own logic to figure out how we 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 [53], and evolutionary learning, where they account for the biological evolution and that it can be seen as a learning process.

In this thesis we only consider supervised learning. 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 respons variable can either be qualitative or quantitative in nature. For the qualitative respons 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 $x = 0.5$ is not in the data points given so it is needed a way to *predict* the value. 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 respons 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 need to pick a value in between our data points and use our functions h and g to predict its values and compere the result to see which is better. mby connect it tighter with the handwritten example? This does not seems as very intelligent behavior, but the problems of interpolation can be very difficult in higher dimensional space. This will 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.

In this work we want to predict characteristics of batteries , meaning that our task is a regression problem.

3.1.4 Data collection, Preparation, Features and Feature Selection

Normally the data collection is a enormous 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 different 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 before doing a complete analysis.

A common problem is that there is too much 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 thing 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 our 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 will be handled later(Ref).

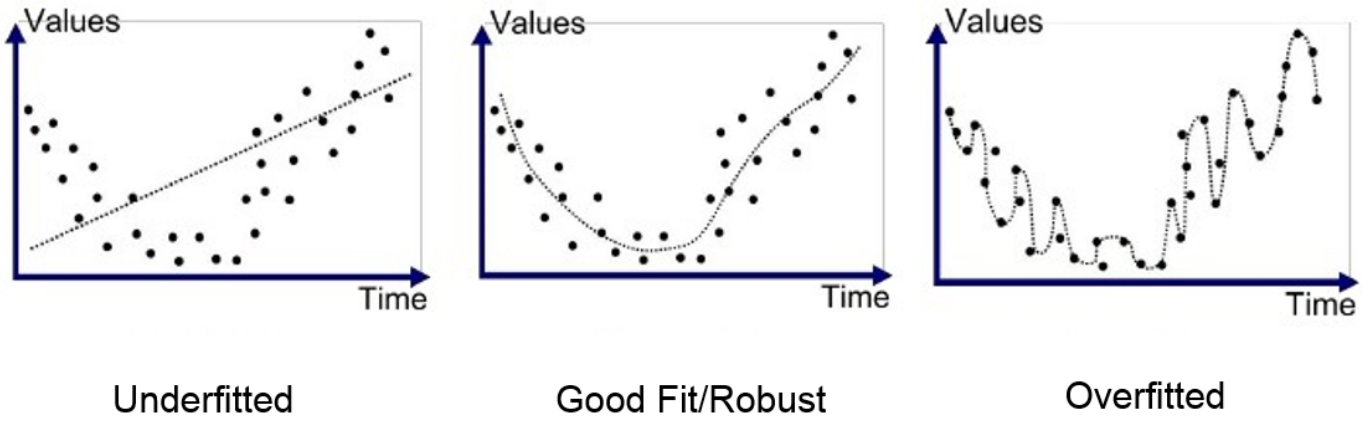


Figure 6: Simplified illustration showing the concepts of bias-variance problem. Left to right; high bias, low bias and low variance, high variance 6

.

Our machine is bias if it generalizes to 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 6.

Variance is the error that stems from high variability, and the number of degrees of variability in most machine learning algorithms is huge([53]). In simple terms there is a low degree of generalization. It might be a perfect fit, but as soon as new data is introduces our predictions plummet. This is commonly referred to as over-fitting, see right on figure 6.

A good way to understand the idea of bias-variance tradeoff is that 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*[54]. 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.

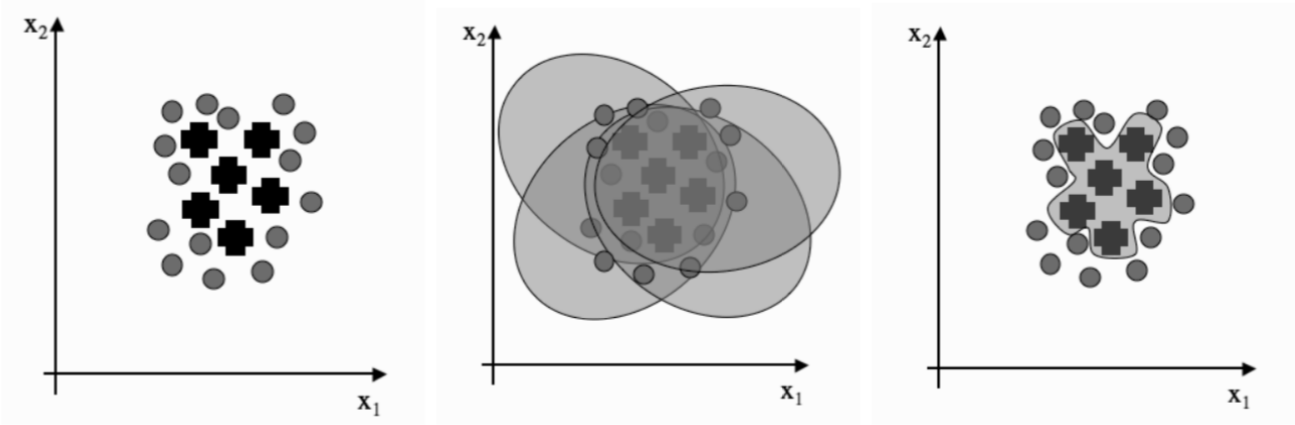


Figure 7: Combining a lot of different classifiers trained on the same data, which in combination can make a much better decision boundary on the target data. Adopted from [53]

Ensemble methods are particularly usefull in machine learning when there is little data, as well as when there is much data, this is heavily due to cross-validation, see(Ref to other section on cross-validation).

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 datapoints.[53].

Decision trees are structured much like a regular tree8; 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.Example? If overcast, play. Rain -> more questions 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. An advantage of random forest being based in decision trees is that the algorithm

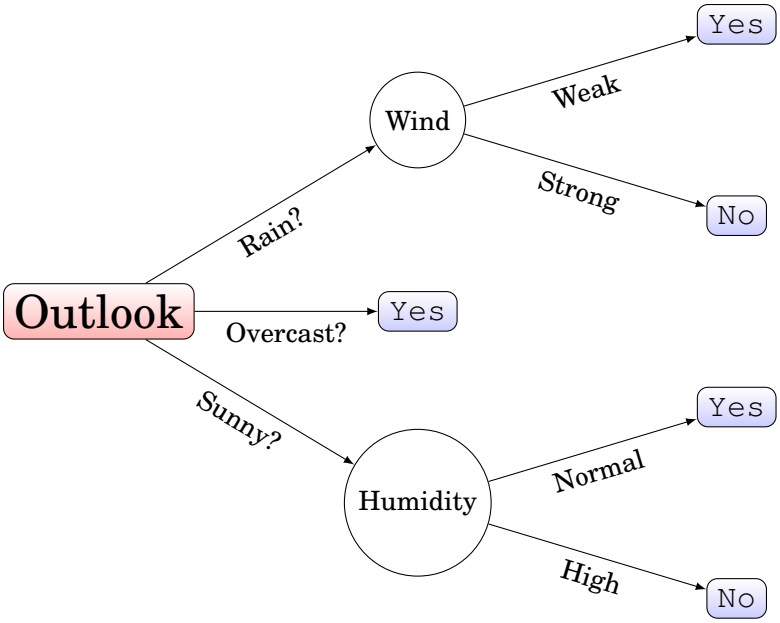


Figure 8: A simple example of a decision tree for playing tennis. Root in red, leaf node in blue. Adapted from [55]

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.

3.3.3 Random forest

Random forest is ensamble such a 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 the bootstrapping in random forest, one sample is taken from the dataset for each tree.

Then a new parameter is introduced, at each node a random subset of features are given to the tree, and 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*, that is, reducing the complexity of decision tree by removing the parts of the tree that does not help the classifier. This 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 point by point algorithm, see the appendix(REF to appendix).

The reason for not using cross-validation in the learning algorithm, which is common in other machine learning methods (Ref to cross-val), 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 [56] and the findings from both our group [**fanourgakis2020automatedML**] and Shandiz and colleagues [22], that clearly states that random forest is the go to machine learning algorithm when you are not sure where to start. Another reason and another of the main advantages of RF is that it is fast and does not require any particular optimization of its hyper-parameters (i.e. number of decision trees). On the other hand, methods like support vector regression (SVR) requires an extensive search for the optimum hyper-parameters before providing reasonable results.

3.4 Evaluation method

3.4.1 Mean square error

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

$$\text{MSE}(\epsilon) = \frac{1}{n}\sum_{n}^{n-1}\epsilon^2 = \frac{1}{n_{\text{samples}}}\sum_{n}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2 \tag{3.1}$$

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.

3.4.2 Root mean square deviation

The Root mean square deviation, or root mean square Error (RSME), is the squared for 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.

3.4.3 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:

$$\text{MAE} = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i| \tag{3.2}$$

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.

3.4.4 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.

$$\text{WAPE} = \frac{\text{MAE}}{\bar{y}_i} \times 100 \tag{3.3}$$

3.4.5 R^2 score - The Coefficient of Determination

In regression validation the R^2 is the standard when it comes to measuring goodness of fit[57]. In straight terms it is the proportion of the variance in the dependent variable that is predictable from the independent variable (S).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{3.4}$$

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, than 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.

3.4.6 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 [57].The approach for the technique is simple. Instead of doing a regression on the entire data set, it is first segmented into k number of subsets 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 'control' or 'validation' set while the rest of the subsets 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, for example 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. After all this is done 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.

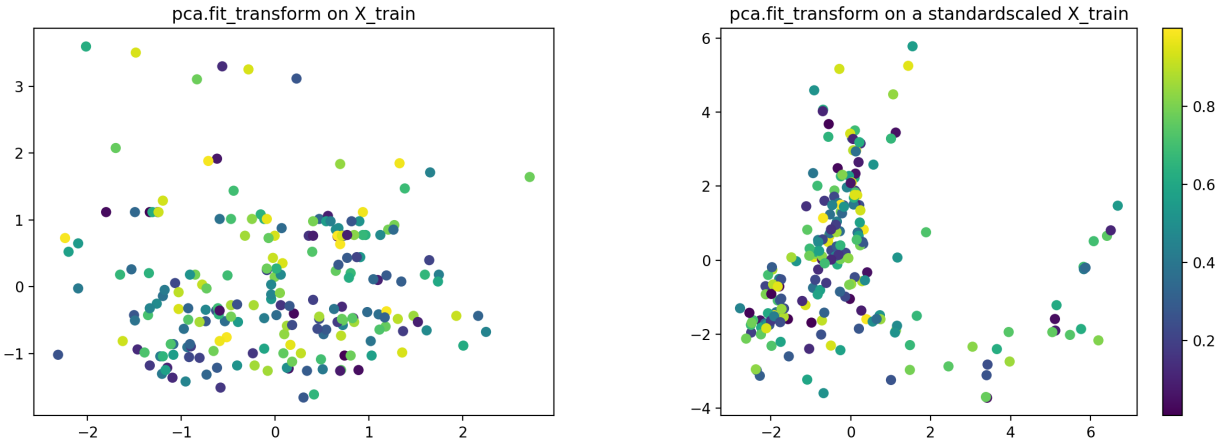
Cross validation techniques are extremely useful when the gathering of new data is difficult or, sometimes, even impossible, as we are using the extra computational power at our disposal to squeeze the most amount of relevant information out of our data.

FIGURE? link to good crossvalidation. `https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation`

3.5 Principle Component Analysis

Principle Component Analysis (PCA)[57] is a procedure that uses orthogonal linear transformation to reduces the amount of feature subspaces. It goes under different names in different fields, but the most recognizable might be Singel Value Decomposition. This is done by converting a set of possible correlated variables into a set of uncorrelated variables, called principle components (PC).

The PC are arrange so that the first PC has the largest variance, meaning that it accounts for as much of the variability in the data as possible. The second PC does the same, it accounts for as much of the variability as possible with the constraint that it is orthogonal to all the former components. These orthogonal vectors are linear combinations being an uncorrelated orthogonal basis set. Graphically the shortest vectors effects the predictions the least. PCA is sensitive to the relative scaling of the original variables, so in *sklearn.decomposition.PCA*, from the library we use, the input data is centered but not scaled, in this step, for each feature before the PCA is done on the date.



(a) (b)

Figure 9: Two scatter plots; First, some of our data before PCA. Second, our data after PCA, showing that there are clearly distinguishable classes.

Part III

Experimental method

4 Method: Better section header

In this section we will introduce the overall approach to the research. First introducing the data set and experimental environment, before going through technicalities in the methods used to represent physical and chemical properties of the electrode materials.

The two most crucial challenges in ML are; Create the dataset, which needs to be big enough and as accurate as possible to better the predictions given the right features. Secondly, finding the right descriptors. If not both of these conditions are fulfilled, then it is not possible to create a reliable ML model. Put in ML part?

All codes are written in Python 3.7 [58] or Fortran98 [59] and can be found here: `https://github.com/sondrt/Creten_Stuff` gitrep private, change?. And all computations are done on a (give computer specifics.), if nothing else is noted Any libraries that uses this? read machine architecture. . The main frameworks and libraries we used for our data mining and data analysis are Python with NumPy [60], JSON [61], Pandas [62], and Scikit-learn [63]. To make the AP-RDF files Fortran90 was used.

4.1 Data set and Experimental Environment

The techniques used in this work are the *volumetric number density*(n), the *Void fraction*(Vf) and the *atomic property weighted radial distribution function*(AP-RDF). Other properties used where introduces in the foundation section 2.4.2. Lastly we quickly mentions the basics of our algorithm.

The first needed for this task was a database(db), we found several, but opted into *materialsproject* [23] [64] (mp), due to their investment in the battery explorer (`www.materialsproject.org/batteries`) [24] [25], which made, the otherwise main task of this work, collecting the data, much easier, and a functioning API [65]. The db has a sizable amount of information on electrodes available. Mp has 16128 conversion electrode and 4401 intercalation electrodes. We decided to limit the project to the intercalation electrodes, due to this limiting the variation of materials in the db.

First of all mp had the reduced cell formula with consistent CIF files for all voltage pairs, that is; both the charged and discharged material. Secondly many different characteristics or voltage pair properties are present. Some of these characteristics, that we used as targets in this work, are; Average voltage, Gravimetric and volumetric Capacity, Specific Energy (Wh/kg), Energy density (Wh/l), and a measurement of the stability; energy above hull measured in

$\mathrm{eV/atom}$. Other properties that where in the mp db and to some extent tested as predictors for each material, both charged, and discharged are; the space group, energy per atom, volume of the unit cell, volume change in percentage, band gap, density, total magnetization, number of sites, and elasticity.

The database contains more than 4400 intercalation electrodes, where we have used 2291 Lithium-ion batteries, and 360 Magnesium-ion batteries, for our analysis. This could easily be expanded to the 321 Natrium-ion batteries, and 481 Calcium-ion batteries in the db. With new compounds being added to the database continuously, including many new structural predictors, there is a high likelihood of an increase in accuracy over time, due to the db growing. Our method were tested on Mg-intercalation electrodes as well as Li-intercalation electrodes, then compared to each other to identify correlations on the valuableness of our predictors.

It is important to note that we have a minimum of two predictor per property of the material at any given run. This is due to how we defined each battery. They have at least one charged and one discharged state, and only one target(ref to theory machine learning). For any given property we have one value calculated for the charged material, and one for the discharged material.This means that we predict for a specific charged- or discharged half cell configuration.

Exemplified by the battery $\mathrm{Mg(CrS_2)_2}$ with the battery ID: $mvc-1200000091$, it has two material ID's, one for the discharged- ($mvc-91$), and one for the charged($mvc-14769$) material. This pair will be referred to in this section.

4.2 Volumetric number density

Volumetric number density, n, is used to describe concentration of countable objects. And is defined as:

$$n = \frac{\#\text{ of atoms}}{\text{Volume of the unit cell}} \tag{4.1}$$

Where *Volume* is the volume of the unit cell. write volume of unit cell in formula?

Technically, in the volumetric number density, there is a predictor for each individual element. That is; if the intercalation battery framework is $\mathrm{Mg(CrS_2)_2}$ then the the number density for; magnesium, chromium, and oxygen, related only to that material, will be predictors. The charged material will have the predictors with the values: $S_{vol} = 36.6292$ and $Cr_{vol} = 18.3150$. While the discharged material will have the predictors with the values: $S_{vol-dis} = 30.1286$, $Cr_{vol-dis} = 18.3150$, and $Mg_{vol-dis} = 7.5321$. All other elements still exists 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. This quality is uniq for the volumetric number density, and all other predictors minimize "empty" columns.

It is probable that such a direct measurement of a geometrical aspect could be a good predictor due to the physical significants of the information. If RF were applied on to the entirety of the CIF file, it is probable that it would make a bad fit, due to the bias-Variance-trade-off as

mentioned in section 3.2, and because of the difference in complexity of some of these files. ref to two vastly different files in the appendix or on github?

4.3 Void Fraction

rewrite

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. Poreblazer [66]. We measure the accessible void, that is, the total amount of free space in the material and it is a quantity that can be measured experimentally.

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. and the pore volume (v_{pore}) and the porosity (θ) can be computed from:

$$v_{pore} = \frac{n_{N_2}^{ads,satd}}{\rho_{N_2}^{liq}} \tag{4.2}$$

$$\theta = v_{pore} \cdot \rho_{cryst} \tag{4.3}$$

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 exists pores that are non accessible for a nitrogen atom, so the experimental data using nitrogen is different form the DFT calculations using i.e. helium.

There are many different computational method to obtain the pore volume where each one computes slightly different portions of the full volume. We have here used two approaches for measuring the pore volume. Frist, the geometric pore volume,Ge_{pv}, which is defined as all the free volume of the unit cell. Secondly helium pore volume ,He_{pv}, where the unit cell is probed with realistic intermolecular potential is tested. 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 in the supportive information/github.

Void Fraction is a characterization method for microporous crystals and have had great success in metal organic frameworks (MOFS), as demonstrated also by the team of supervisors.

In case of dens materials like the one we consider in this work, the void fraction should not be a good predictor. However we decided to include it in our tests in case the space occupied by the ion in the discharged material would impact our prediction, as will be discussed later. (REF)rewrite.

UFF [67]

4.4 AP-RDF Descriptors of Electrode materials

Atomic property weighted radial distribution function (AP-RDF)**??** was found to be a good predictor which also, when tested by the PCA(REF to theory part), exhibited good discrimination of geometrical and other properties, in one of their cases, gas uptake.

One of the methods found, that seemed to yield good predictions dependent on chemical properties where the Atomic property weighted radial distribution function, successfully used on MOFS. **??** Due to it looking reasonable we decided to try it out.

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.(REF) 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)Do I need a ref here? and the RDF scores will be uniquely defined inside of the unit cell, per material-ID. The RDF can be expressed as:

$$RDF^{P}(R) = f \sum_{i,j}^{\text{all atom pairs}} P_i P_j e^{-B(r_{ij}-R)^2} \tag{4.4}$$

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 above. 4.4

Summing 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 in the appendix with an operational pdf.(REF)

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.**??**. We confirmed their findings .. though with drawback related to the size... which are flawed in our case. In our opinion, we think that this is a fundamental drawback, and the results depends on the size of the simulation cell(which can be made by replicating the unit cell).

INSERT BILDET AV PLOT AV AP-RDF.

4.5 Algorithm

First of all, a wrote a program for "scraping" the *materialsproject*webpage for batteries(0). This gave us the possibility to gather all the available resources on the batteries in the database in a fast and effective manor, as well as updating these CSV files of battery-IDs. (ref)

We then run a second program that downloads all the information on the materials that matches a material-ID correlated to a battery-ID(1,2). (ref) Before constructing a CIF file structured so that all the battery-IDs, charged-material-IDs, and dischargerd-material-IDs are correlated with the information on the charged and discharged properties.

After, the volumetric density fraction is calculated(3) and added to the main CSV file for both charged and discharged materials. While the CIF files are being processed for Poreblazer(4) where the void fraction is calculated(5,6).

Then we merge all our CSV files based on what properties that we are interested in and makes a CSV file called for_ML.csv(7,8) that we feed into our random forest algorithm(9). We then run cross validation, MSE, and plot what we are interested in(10).

In addition we also tested for different machine learning algorithms, as mentioned(ref), but these were only to test the reliability of our model, and will be discussed in the discussion section(ref)

maybe add this in the appendix?

Algorithm:

Steps for use of python scripts:

mp_battery_scraper.py

0: Scrape batteries with a given working ion from the Materials Project battery explorer (https://www.materialsproject.org/#search/batteries)

fillproperties.py

1: Download all materials that match a material_id correlated to a battid.
Output files: directory cif_info_dir/<material_id>_prop.dat

add_features.py

2: Gets and adds the material specific features from the JSON dump to a csv.
Output files: material_properties.csv

elements.py

3: Calculate the density fractions for all materials.
Output files: out_csv_dis.csv

forPoreblazer.py

4: Download the CIF files as JSON for all materials correlated to a battid.
Output files: directory cif_for_poreblazer/<material_id>_cif.dat

process_cif.py

5: Extract the CIF information from the previous JSON data.
Output files: directory cif_for_poreblazer/cif_files/<material_id>_cif.dat.csv

process_cif.py

6: Extract void fraction with poreblazer using the CIF files.
Output files: helvol_geomvol.csv

merger.py

7: Merge charged and discharged for all properties
Output files: allFiles.csv

prep_csv.py

8: Select predictors and targets for ML

```
        Output files: for_ML.csv

        randomforest.py
9:  Run randomforrest
        Output files: Depending on what being saved: ./Results/*

        crossvalidation.py
10: Run cross-validation, remove outliers.

11: ???

12: Profit!
```

Part IV

Result & Discussion

5 Result section title

The accuracy of the ML method in this work is evaluated using R-squared (R^2), "Mean" of the training target values (y), 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 indicates a better ML predictions. Before each run a principal component analysis were done and the singular values that collectively accounted for 99% of the variation were selected for the RF run, the last 1% of the variation, (i.e. the other predictors) were removed.

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

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

The targets they are tested on are; Average Voltage (AV), Gravimetric Capacity (GC), Volumetric Capacity (VC), Specific Energy (SE), and Energy Density (ED). After all sets of descriptors were tested, they were combined to make the best possible classifier.

The two data sets are chosen much due to their size. The Mg-ion db has 356 different batteries, while the Li-ion db has 2081 different Li batteries, more than five times the size of the Mg-ion db. Both of these databases are relative small form a ML perspective, but are big enough to show correlations. The impact the difference in size has on the predictions accuracy will be examined.

5.1 Material specific properties

The "material specific properties" (msp), the properties related to every individual material, both charged and discharged, where tested first. 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 **??**. The Materials Project database included

"formation energy per atom" for most structures, but due to "None" values and no significant change in our predictions, probably due to no new variance, we decided to leave this predictor out of the result section. The results for msp can be seen in table 1 and in table 2.

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

From Table 1, it is clear that msp are correlated to the targets, with especially high values for AV. The accuracy falls and uncertainty for the predictions rises in capacity related targets, and for SE and ED. The WAPE also indicates that the uncertainty is high for SE and ED. It is obvious that we either lack part/s of the picture, due to R^2-train score being so low (not closer to 100%, which shows that the ML algorithm can not make a perfect regressor even when it knows the answers). Or that our data contains to much noise. It is likely to be the case for all our runs, because we do not know what predictors are needed. The msp seems to be a considerable part of the puzzle and will be used in later combinations of predictors.

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

When applying the same predictors on the Li- battery db we observe a fall in prediction accuracy on all targets except VC, but the uncertainty of our predictions rises for both GC and VC. For AV SE and ED our weighted absolute percentage error sinkes considerably, and thus the results are more likely to indicate the correlation between our predictor and the target more accurately. The increase in data seems to not effect our predictions to a large extent. The R^2-train is increased by a negligible amount.

5.2 Volumetric number density

Mg-ion framework

Volumetric number density (vnd) as described (4.2) are shown, first for the Mg-database than 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 test, 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

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

There are a couple of different results that are particularly interesting. First of all; The use of all vnd-predictors does not yield the best predictions in all cases. There are targets that

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

respond better to the use of one state of material, charged or discharged, than the other. Gravimetric and volumetric capacity responds particularly well to the discharged materials only, but improve (by a small percentage $3-5\%$) when given the combination of materials. AV are approximately the same (change of 0.7%), while SE and ED improve by 9%. This seems reasonable due to the discharged materials containing more information than the charged materials (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).

As expected, the combination of both the charged and discharged materials give the best results, and as can be seen from the number of components used, a lot of information overlapp between the discharged and charged materials.

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, and is definitely a part of the physical picture. Noticeably the R^2-train is also considerably better (than for msp) when including the specific atoms as we do in vnd.

As expected the 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 on AV 17/21 components where used to account for 99% of the variance in the data, for the discharged materials 19/30 components where used, while when combined only 26 out of the total 51 components were necessary to achieve the same 99%. This indicates a information overlap between the predictors.

Lastly, the quality of our estimator drops significantly 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?

Predictions on Li-ion intercalation frameworks with vnd

The same technique, as used in the previous section, was applied to the mp Li-ion db, with 2108 battery frameworks instead of 365, as were the case for Mg-ion intercalation type db. This is a bigger database with different and more unique atoms, which is likely to calls for more variability and 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, as were with the Mg db. The charged materials seems to be a better predictor for energy density, and average voltage. While the discharged materials are better at predicting capacity. When predicting the specific energy they seems to be equal in their predictions. 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.

Average voltage predictions are 56% with high certainty, lower than the predictions for the Mg-ion db, but a bit higher accuracy. The gravimetric capacity and specific energy are lower in the Li-ion db, with higher accuracy. While energy density is approximately equal and equally accurate.

Onward, both the charged and discharged group of predictors, for both db, will be used in further predictions.

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

5.3 Void fraction

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

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

As a predictor for the magnesium db, 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 very close 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.

On the Li-ion db the same trend as on Mg-ion db seems persistent. More data seems to emphasize that the void fraction is not a good descriptor for the given targets. The best predictions are still on VC, followed by GC, but the scores are not high enough to indicate a proper correlation.

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

On a general note, 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, and yet it is only possible 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 thing, 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.

5.4 AP-RDF cross-product approach

The AP-RDF was first tested with a cross-product approach, this is explained in the method section **??**. As is clear form the table (11) the first approach to AP-RDF did not give a good correlation with the targets and the error is to big. The idea was that it is better for a RF approach to consider many rows rather than many more columns.

Table 11: 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

5.4.1 Row approach to AP-RDF

The second approach gave a better correlation with the targets, here we added a new columns per value from the RDF, as explained in the method section (**??**).

This approach seems more promising with high prediction values for several targets (GC, VC, SE, ED), and relatively low uncertainty. The correlation is stable when we switch to a db of a bigger size, 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 12: 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 13: 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

5.5 Combining predictors

When combining the predictors; msp and vnd, for the Mg db the predictions increase for AV and SE (3%, 7%), while it stays relatively still for GC, VC and ED (change $< 2\%$). The weighted absolute percentage error falls which indicates that the predictions are more reliable.

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

When combining msp and vnd, for the Li db, the trends here are consistent with the once from the Mg db. For AV and SE the predictions increases (13%, and 11%). For the predictions in general, the combination of predictors either heightens the predictive score or lowers the error.

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

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 of several possible reasons might be that the stability of a material is connected to that particular material, and using a combination of both charged and discharged properties can confuse the ML algorithm due to the amount of unnecessary

information. The predictions on charged and discharged stability can be found it the 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.

5.6 Combining target and predictors

In a last effort to evaluate the method some of the targets were introduced as predictors. The idea being that if one of the targets is simple to compute through 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 db will be considered, then the Li 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

As a final note: If 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.

Part V

Summary

6 Conclution and future work

6.1 Batteries

6.2 future work

6.2.1 improving method

References

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

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

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

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

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

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

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

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

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

[47] Mahesh Datt Bhatt and Colm O'Dwyer. "Recent progress in theoretical and computational investigations of Li-ion battery materials and electrolytes". In: *Physical Chemistry Chemical Physics* 17.7 (2015), pp. 4799–4844.

[48] 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.

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

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

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

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

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

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

[55] *Example: Decision treeStefan Kottwits*. `http://www.texample.net/tikz/examples/decision-tree/`. Accessed: 2020-01-23.

[56] 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.

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

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

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

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

[61] 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.

[62] 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.

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

[64] 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.

[65] 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 (Feb. 2015), pp. 209–215. DOI: 10.1016/j.commatsci.2014.10.037. URL: http://dx.doi.org/10.1016/j.commatsci.2014.10.037.

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

[67] 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.