

UiO : **Department of Physics**
University of Oslo

Sondres master utkast

26.02.2020

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

I	Introduction	1
1	Introduction and Overview	1
1.1	Motivation	1
1.2	History and evolution of batteries	1
1.3	Lithium based batteries	4
1.4	Magnesium batteries	6
1.5	Earlier work	6
1.6	Scope of the thesis	6
1.6.1	Research Question	7
1.6.2	Approach	7
1.7	Structure of the thesis	8
II	Foundations	9
2	Batteries	9
2.1	Cell operation principles and design	9
2.1.1	The manufacturing process of an Alkaline battery	10
2.1.2	Theoretical Cell voltage, Capacity, specific Energy, Energy density	11
2.2	Cell limitations & definitions	13
2.2.1	Polarization	13
2.3	Battery chemistries	13
2.4	Intercalation batteries and why Li	13
2.5	Electrodes and features	14
2.5.1	Properties of materials	15

2.6	On Mg- and Li- batteries and their state today	17
3	Machine Learning	18
3.1	The basics of Machine Learning	18
3.1.1	Example time!	18
3.1.2	Supervised and Unsupervised Learning	19
3.1.3	Regression and Classification Problems	19
3.1.4	Data collection, Preparation, Features and Feature Selection	20
3.2	Bias-variance tradeoff	20
3.3	Random Forest	21
3.3.1	Ensemble learning	21
3.3.2	Decision tree	22
3.3.3	Random forest	22
3.4	Mean square error and Root mean square deviation	24
3.5	R^2 score - The Coefficient of Variation	24
3.6	Principle Component Analysis	25
3.7	K-fold cross validation	25
3.8	Mutual Information**	25
3.9	Machine Learning \times batteries	26
III	Method	27
4	Some Method title	27
4.1	Data set and Experimental Environment	27
4.2	Volumetric number density	28
4.3	Void Fraction	28
4.4	AP-RDF Descriptors of Electrode materials	29

4.5	Algorithm	30
-----	---------------------	----

IV Result & Discussion 33

5 Result section title 33

5.1	Random factors from database.	33
5.1.1	Average Voltage	33
5.1.2	Capacity	33
5.1.3	Energy Density	33
5.2	Volumetric number density	33
5.2.1	Average Voltage	34
5.2.2	Capacity	35
5.2.3	Energy Density	35
5.3	Void fraction	35
5.4	AP-RDF	35
5.5	Stability	35
5.6	Geometrical descriptors	36

V Summary 37

6 Summary and future work 37

6.1	Batteries	37
6.2	future work	37
6.2.1	improving method	37

List of Figures

1	A voltaic pile, the first battery [18]	2
---	--	---

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. Figure recovered from [19]	3
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. Figure attained from materialsproject [materialsprojectTiS2].	5
4	Schematic illustration of the first Li-ion battery $\text{LiCoO}_2/\text{Li}^+$ electrolyte/graphite), Figure acquired from [goodenough2013li].	6
5	Simplified illustration showing the concepts of bias-variance problem. Left to right; high bias, low bias and low variance, high variance. Adopted from [17]	21
6	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 [10] . . .	22
7	A simple example of a decision tree for playing tennis. Root in red, leaf node in blue. Adapted from [6]	23
8	Mg- prediction accuracy results 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(REF). . .	34
9	Li- prediction accuracy results 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(REF). . .	34
10	Mg- prediction accuracy results 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(REF). . .	35
11	Mg- prediction accuracy results 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(REF). With the AP-RDF as predictor.	35

List of Tables

Part I

Introduction

1 Introduction and Overview

Introduce the task at hand.

In this chapter the background and motivation for this paper is presented. The historical development of batteries are presented shortly and former work are presented.

1.1 Motivation

The motivation for this work originates from the rapidly increasing demand for better batteries, both for vehicular and stationary applications. Better batteries mean better products, 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

1.2 History and evolution of batteries

Never been a more needing market for batteries. 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 in the year 1800 by Alessandro Volta (1745 - 1827) that made the first known battery[16]. His invention was the voltaic pile, zinc and copper plates stacked on top of each other with sheets of bine-soaked cardboard between each plate. The revolutionary property of the voltaic pile was that it could produce a stable current over some time, and not just short sparks of electricity. This invention was the foundation of todays modern battery. 1



Figure 1: A voltaic pile, the first battery [18]

It took almost 40 years, 1836, before the British inventor John Frederic Daniell Voltas continuation with the discovery of the Daniell cell [5]. The Daniell cell is constructed with to 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 V through the reaction shown below, 1.1.



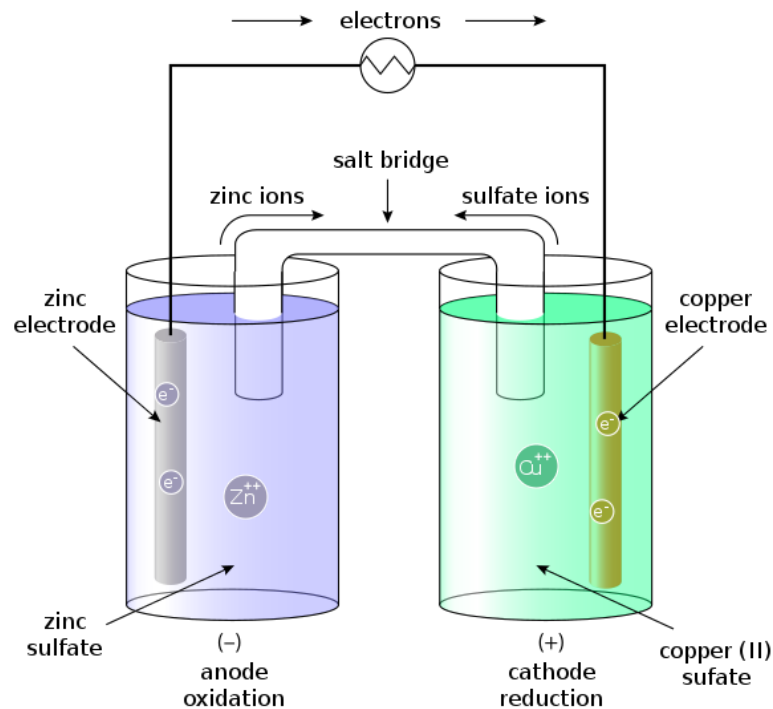
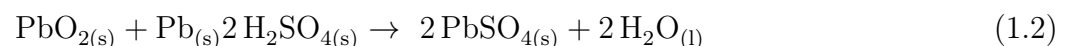


Figure 2: A draft of a Daniell cell. The anode is a piece of zinc and the cathode a piece of copper. The salt bridge transports ions between the solutions and the electrons moves through an external circuit. Figure recovered from [19]

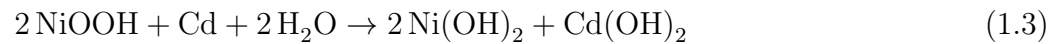
In 1859 constructed the french physicist Gaston Planté the first lead-acid battery. The battery could be charged by applying an external opposite potential, and it was the first secondary battery every made. Planté rolled two lead plates into a spiral, separated by rubber strips, so that the plates would not touch. The lead-acid battery was special due to the electrolyte being a active part of the chemical reaction in the battery. The electrodes in the battery were lead anode, and lead(IV)oxide cathode, immerse in sulfuric acid. The overall reaction is shown beneath: 1.2. Both the anode and the cathode are made into lead(II)sulfate during discharged. The charge is depleted in the electrolyte when the battery is completely discharged(The sulfuric acid has a lower density). Charging changes the electrolyte back into concentrated sulfuric acid.



The open circuit voltage(V_{OC}) for a lead-acid battery are approximately 2 V. It is custom to attach these batteries in series to attain a higher voltage, typical 6 or 12 V. These batteries have a shelf- and cycle life. Normally more than 10 years or 1000-2000 cycles. They are still in use in modern cars. Lead-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. Therefore, one of the many goals of battery producers are to replace lead-acid batteries with better alternatives.

Nickel-cadmium(NiCd) batteries was first described by the swede Waldemar Jungner in 1899 [4].

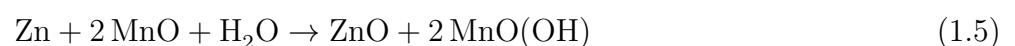
These batteries rose fast in popularity due to their high energy density, low weight, long shelf life, and their relative fast recharge. Typically they yield a voltage of 1.4 V. The cathode is made of Nickel oxide hydroxide, and the anode is made of metallic cadmium, while the alkaline electrolyte is a basic solution of potassium hydroxide. The Specific energy of a typical Nickel-cadmium is 40 – 60 Wh/kg. Equation 1.3 shows the overall reaction of such a battery.



Nickel-metal hydride (NiMH) batteries was first commercialized in the 1980's and had several similarities with the NiCd batteries. The main difference is the anode where in NiCd batteries it is Cd, while in NiMH it is replaced by an alloy that makes metal hydrides (MH). NiMH batteries has the same electrolyte as NiCd batteries, a watery solution of potassium hydroxide. The nominal cell voltage of such a battery is typically around 1.2 V and the Specific energy is 60 – 120 Wh/kg. Equation 1.4 shows the overall reaction of a NiMH battery.



A primary cell is a battery that can not be recharged after use. 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. Anodes of zinc, cathodes of manganese oxide and a electrolyte of potassium hydroxide, from where it gets its nickname. A typical alkaline battery delivers a nominal cell voltage of 1.5 V. The overall reaction for a alkaline battery is shown below 1.5.



1.3 Lithium based batteries

In 1975 discovered Micheal Stanley Whittingham the intercalation electrodes for lithium and other alkaline metals [**whittingham1975lithium**]. 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 can easily move in and out of the structure 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 moves out of the TiS_2 -structure and titanium oxides back to titanium(IV) again. This discovery was the start of a major research investment in cathode materials of sulfite and other chalcogens in the 70-80's.

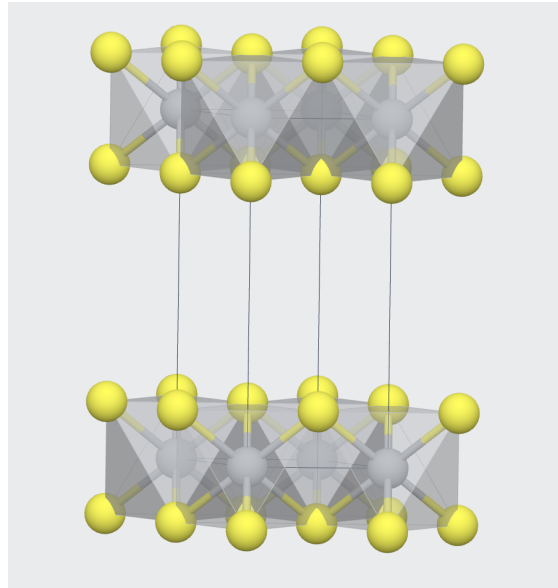


Figure 3: The two-dimensional trigonal omega-like structure of TiS_2 . Observing from a slight angle along the b-axis. The titanium in grey, are occupying the octahedral holes of sulfur, in yellow. Lithium-ions would intercalate into the space between the TiS_2 layers. Figure attained from materialsproject [materialsprojectTiS2].

It is of the utmost importans that the Lithium-ions in a lithium battery can easily be intercalated and extracted (Can I phrase it like this?) in a reversible manor, and such layered structures are especially good for this purpose. In 1980 introduced John B. Goodenough LiCoO_2 as the cathode material for lithium batteries, a material that gave a V_{OC} of 4 – 5 V, which opened up many possibilities to alternate the negative electrodes. And they found a current density of up to 4 mAcm^{-2} [mizushima1980lixcoo2] [goodenough1980solid]. Even though the properties where exceptionally good at the time, the batteries were still not commercialized due to metallic lithium being unstable, and deemed to be an unsafe anode. This was due to dendrites growing out of the anode that short circuited the battery.

In 1990 introduces Sony lithium batteries on the commercial market. They had exchanged the metallic lithium anode for graphite which reduced the growth of dendrites on the anode. The electrolyte was a 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. While lithium batteries has a anode of metallic lithium. This nomenclature is transferable to other type of batteries like, magnesium/magnesium-ion batteries.

Figure 4 shows a draft of a lithium-ion battery with LiCoO_2 as the cathode and graphite as the anode. During discharged the lithium-ions moves from the anode, through the electrolyte and separator to the cathode. The electron moves from the anode to the cathode through a separat circuit, where the electrical energy can be extracted. The overall reaction is shown in equation: 1.6

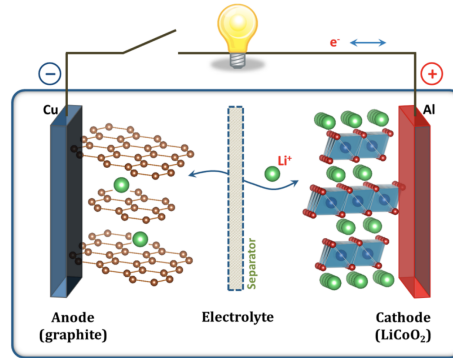
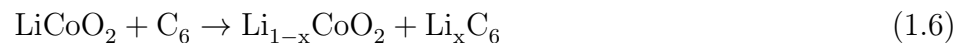


Figure 4: Schematic illustration of the first Li-ion battery $\text{LiCoO}_2/\text{Li}^+$ electrolyte/graphite), Figure acquired from [goodenough2013li].



The cathode materials used in

1.4 Magnesium batteries

1.5 Earlier work

1.6 Scope of the thesis

Write more about: PV, wind, solar -> storage.

"At what cost can one charge stationary secondary batteries with solar/wind power efficiently, and provide a centralized energy storage? For what shelf and cycle life? with how rapid a response a power outage or fluctuation in the grid. And with how large a capacity, at a competitive cost? " **where?**

Batteries are vastly complex and much efforts have been devoted to the development of these. Yet, with all the efforts put in to electrochemical 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[1], 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 proposes a methodology to predict these properties accurately without the need of big scale simulations, or computer heavy calculations. Using state of the art machine learning, and base

properties of all ready existing databases, we propose a set of predictors to see if we can 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. And it is of the general public's interest to improve several of these characteristics, both for once pleasure (i.e. several more hours to play candy crush), and the people's general safety (i.e. electric car crashes, and Samsung phones). In this work we will focus on the base properties; voltage, energy density, specific energy, and the physical stability of the materials. No real battery fulfills all of these requirements. And most real batteries underperform. We therefore leave the safety issues to the experimentalists.

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

1.6.1 Research Question

How to better batteries?

RQ1: Is there potential for the use machine learning to ease 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?

1.6.2 Approach

The choice of features examined in this work is inspired by an extensive survey done on similar project especially in the field of Metal Organic Framework(MOFS) done by my supervisors in Crete, and dictated, to some degree by the lack of more data.(REF)

To answer the questions

In order to evaluate the effect of different features, a prediction approach using principle component analysis was utilized. First we used physical descriptors, geometrical properties of the unit cell, for instance. This because it was greatly efficient in similar studies(REF) on MOFS, and is straightforward. Other descriptors are needed, and the void fraction seemed like the next obvious one, that is both geometrical and easy to obtain. After

1.7 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 explained before rounding up our results so far before trying to put this all into perspective.

Part II

Foundations

2 Batteries

The basis of batteries needs an introduction. In this section, batteries and how they are designed and how they operate will be explained. A special focus will be dedicated to electrodes which are the part of the battery that this work is focused on. Some of the more essential properties related to the chemistry of the work will also be introduced.

"The most advantageous combination of cathode and anode materials are those that will be lightest and give a high cell voltage and capacity. [14]"

Some of the most important properties in an anode are; High coulombic output(Ah/g), good conductivity, stability.

" Practically anodes are metals. Li- are a good example of this, it is the lightest metal with a high value of electrochemical equivalence. With the development of intercalation electrodes, lithiated carbons are finding wide use. Lithium alloys are also being explored for use as anodes in lithium-ion battery.

The cathode must be a good oxidizing agent, be stable, when in contact with an electrolyte, and have a useful working voltage.

2.1 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 composed 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 an oxide is iron oxide, this oxide is very fragile, but some are oxides that are strong enough to be shaped before used in the cell. 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 anode draws oxygen atoms toward it, creating an electrical flow, with electrons moving from the end of the anode to the end of the cathode not immersed in the electrolyte. The electrolyte is typically a liquid, to impart ionic conductivity. It is

often alkaline and when this is the case the battery is a alkaline battery. It can be a solid, but this is less common. The cell will not produce electricity without a circuit that is complete. The electrolyte can, in some designs, 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. [9]

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, as are the one in question (**Not really true, we are only looking at Electrodes.**), it is possible to recharge batteries by reversing this process by applying an external electrical power source, so that it creates a 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. **Should cover these shortly in another subsection? Write something about Magnesium and Li- as a minimum.**

2.1.1 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, in a Alkaline battery like this, 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 designs, 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 take 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?

2.1.2 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:



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



with the overall reaction being:



Whenever there is a reaction, there is a decrease in the free energy of the system, namely standard Gibbs energy, and is defined as:

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

Where n is the number of electrons in the reaction, and F is the Faraday constant.

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 from the standard electrode potential.

$$\text{Anode} + \text{cathode} = \text{standard potential}$$

example?

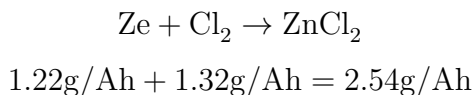
In the situation where we are not in the standard state, the voltage E of a cell is given by the Nernst equation.

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

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

Gibbs free energy of the reaction is the driving force of the battery, and enables it to deliver energy to an external circuit.

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 Zn/Cl₂ cell is 0.394Ah/g



Capacity can be defined as:

$$C = \int I(t) \cdot dt \quad (2.2)$$

And is the i number of electrons or cations exchanged between the negative and positive electrodes. $I(t)$ is the current, i.e. 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 Ah/kg. Theoretically, 1 gram equivalent weights of the active material in grams divided by the number of electrons in the reaction.

The battery can deliver a power that is defined as

$$P(t) = V(t)I(t) \quad (2.3)$$

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

$$W = \int P(t) \cdot dt = \int V(t)I(t) \cdot dt \quad (2.4)$$

Specific 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.

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. *intercalation* 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:

$$\text{Watt-hours/gram} = \text{Voltage} \times \text{Ampere-hours/gram} \quad (2.5)$$

2.2 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.

2.2.1 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.

2.3 Battery chemistries

2.4 Intercalation batteries and why Li

"Among the rechargeable battery technologies (Fig. 11), the current Li-ion battery technology offers the best useable energy and tap density, dominating the worldwide market for mobile electronic devices² for decades. Li-ion batteries have minimal (unwanted) side reactions when a Li ion intercalates into or desorbs from the cathode/anode materials. Their energy efficiency may be further enhanced by lowering the internal resistance of the battery, and they exhibit limited self-discharge, and no memory effects that limit energy density after many cycles. 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 success of 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?

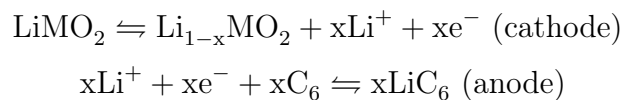
The schematic diagram of the current Li-ion battery based on a carbon based anode (Li_xC_6), cathode (LiCoO_2), liquid electrolyte (LiPF_6 dissolved in a mixture of EC and DMC or equivalent), and separator is shown in Fig. 2.9 In the foreseeable future, Li-ion batteries will be the most practical solution 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

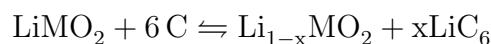
"[2]

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. LiMO_2 where M is a transition metal), through the electrolyte that commonly is Li^+ -containing salt. The reaction, as discussed above 2.1, are:



The overall reaction:



The anode is graphite, thus there is no metallic Li, which makes the Li-ion battery less reactive, therefore 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 electrolyte's 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.

2.5 Electrodes and features

In this section the features used in ML as predictors will be introduced. 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 the temperature, normally 25°C.

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:

$$\text{Anode(oxidation potential)} + \text{cathode (reduction potential)} = \text{standard cell potential} \quad (2.6)$$

The cell voltage is also dependent on other factors including concentration and temperature, as expressed om 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.

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. Some positive value indicates that the material is not stable. While a zero energy above hull indicates that this is the most stable material at its composition.

2.5.1 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 to 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 Mg_3Sb_2 structure would be tested against other Mg_3Sb_2 structures, against Mg and Sb mixtures, and against MgSb and 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 Gaps* are calculated a little different. In general, band gaps are computed with common exchange-correlation functionals such as the LDA [13] and GGA are severely underestimated [12]. Typically the disagreement is reported to be 50% in the literature. Some internal testing by the

Material Project supports these statements; typically, they found that band gaps are underestimated by 40%. 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

T_m

Elasticity

E

Porous Electrodes

In a fuel cell system, the reactant is supplied from the electrolyte phase to 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 having a porous electrode design. A porous design can provide an interfacial area per unit volume that is considerably higher than that of a planar electrode.

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

2.6 On Mg- and Li- batteries and their state today

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 off 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 ..."([11]). At its core the ability to learn by detecting patterns in usually huge amounts of data that, more often than 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([8]). 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 $S = (X_{train1}, Y_{train1}), \dots, (X_{trainN}, Y_{trainN})$ and returns a prediction rule: $h : X \rightarrow Y$. This rule is also called a predictor, in general, a classifier, or a regressor, depending on the problem in question.

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



3.1.2 Supervised and Unsupervised Learning

One of the most basic separations in machine learning is the partition between supervised learning and unsupervised learning. [7]

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-data 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 (marsland2014machine), 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 *response* variable can either be qualitative or quantitative in nature. For the qualitative response variable, let's assume a set of data points \vec{x} and a goal of finding the value of the output y when $x = 0.5$. The value $x = 0.5$ 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 *response* 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, let's 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 *compare* the result to see which is better. *maybe connect it tighter with the handwritten example?* This does not *seem* 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. I underline again that 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(How much should I refer to this here and not later?), 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).

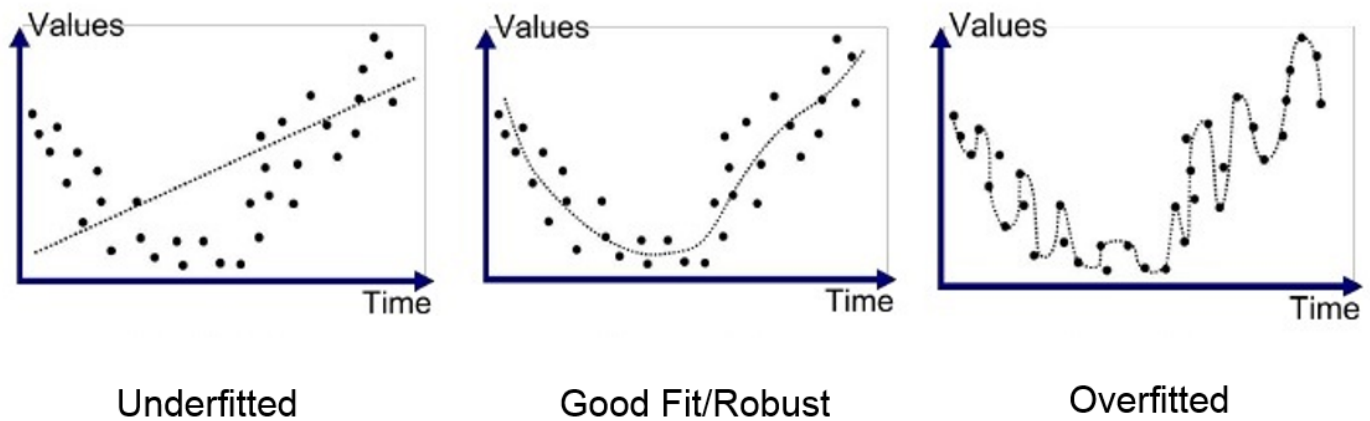


Figure 5: Simplified illustration showing the concepts of bias-variance problem. Left to right; high bias, low bias and low variance, high variance. Adopted from [17]

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 5.

Variance is the error that stems from high variability, and the number of degrees of variability in most machine learning algorithms is huge([10]). In simple terms there is a low degree of generalization. It might be a perfect fit, but as soon as new data is introduced our predictions plummet. This is commonly referred to as over-fitting, see right on figure 5.

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

Should I use more math? I am trying to expaine these things as simple as possible with little mathematical syntax

3.3.1 Ensemble learning

There are many different machine learning algorithms, in this work we have focused on the *ensemble method*; *Random forest*[3]. 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.

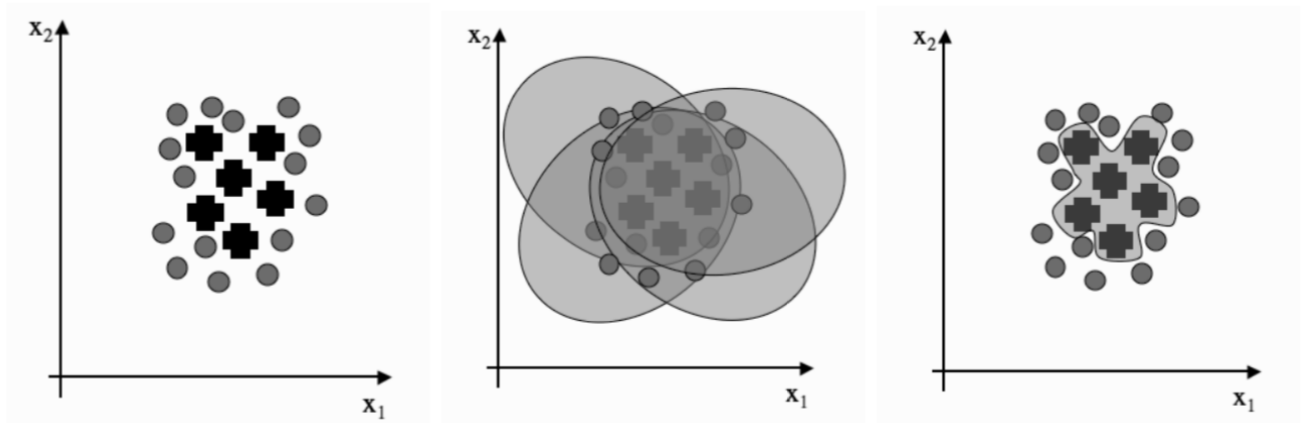


Figure 6: Combining a lot of different classifiers trained on the same data, which in combination can make a much better decision boundary on the target data. Adopted from [10]

Ensemble methods are particularly useful 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.[10].

Decision trees are structured much like a regular tree⁷; 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 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 ensemble 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

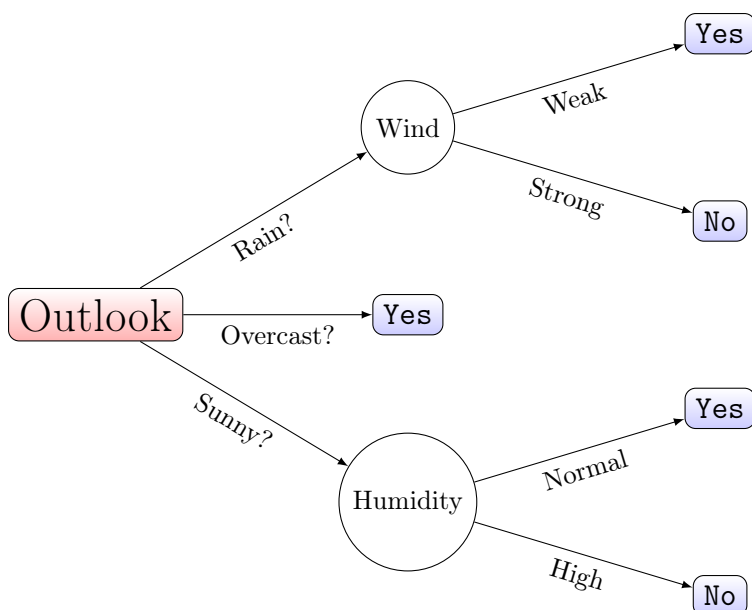


Figure 7: A simple example of a decision tree for playing tennis. Root in red, leaf node in blue. Adapted from [6]

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 that clearly states that random forest is the go to machine learning algorithm when you are not sure where to start.

3.4 Mean square error and Root mean square deviation

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

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

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.

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

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

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 so is a better measure of goodness of fit than a correlation coefficient.

3.5 R^2 score - The Coefficient of Variation

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

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad (3.2)$$

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

If the residual sum of squares (SS_{res}) is low the fit is good. However, this should be compared to the spread of the response variables. After all, if the response variables are all nicely distributed close to the mean, then getting a good SS_{res} is not suspicious. We therefore do a normalization in the fraction, taking the scale of data into consideration. In the simplest polynomial fit, using a zero order polynomial (a constant), our model would just be a constant function of the mean. The sums being equal, returning unity on the fraction and the total R^2 score would be zero. In the other extreme, if the model fits perfectly, then 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.

3.6 Principle Component Analysis

Principle Component Analysis(*PCA*) is a procedure that uses orthogonal linear transformation to reduce the amount of feature subspaces. It goes under different names in different fields, but the most recognizable might be Singular 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 arranged so that the first *PC* has the largest variance, meaning that it accounts for as much of the variability in the data given as possible. The next *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* the input data is centered but not scaled for each feature before the SVD of the data is applied.

3.7 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.(ref=[**cross-valid**]) 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

3.8 Mutual Information**

Ranking of features. This can be useful.

3.9 Machine Learning × batteries

State of the art. What has been done in this field? Similar predictions of the kind we are interested in, have been done by Sandek: [15]



Part III

Method

4 Some Method title

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.

All codes are written in Python 3.7 or Fortran98 and can be found here: (www.github.com/Mahsciencepage ((Give a real REF))). And all computations are done on a (give computer specifics.), if nothing else is noted. The tool we used, for the most part, for our data mining and data analysis is scikit-learn.(REF)

4.1 Data set and Experimental Environment

These are the void fraction, the volumetric number density and the atomic property weighted radial distribution function(AP-RDF). Other properties used where introduces in the foundation section(ref back). Lastly we quickly mentions the basics of our algorithm.

((Sendek(REF) and colleagues were right: They stated that they could not find a sufficient amount of data on ionic conductivity for a proper model to be created, or rather, they only found 40 materials that they used to make their prediction on ionic conductivity, something that our group deemed far to little for our use.))

Needed a database, found *materialsproject*! www.materialsproject.org/batteries gave a database with a good amount of information available. First of all the reduced cell formula with consistent CIF files for all voltage pairs. Secondly many different characteristics or voltage pair properties; Average voltage, Capacity, both gravimetric and volumetric, Specific Energy (Wh/kg), Energy density (Wh/l), and a measurement of the stability ($eV/atom$). Other properties that where in *materialsproject* and to some extent tested to see if they were good predictors are; 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 in more then 4400 intercalation electrodes, where we have used 2291 Lithium-ion batteries, 360 Magnesium-ion batteries, 321 Natrium-ion batteries, and 481 Calcium-ion batteries for our analysis. With new compounds being added to the database continuously, including many new structural predictors. Our method were tested on these different types of intercalation batteries, both with respect to size of database and size of unit cell.

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 to have at least one charged and one discharged state. For any given property, i.e. Volumetric number density, 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.

4.2 Volumetric number density

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

$$n = \frac{\# \text{of atoms}}{\text{Volume}} \quad (4.1)$$

Where *Volume* is the volume of the unit cell.

Technically, in the volumetric number density, there is a predictor for each individual element. That is; if the material is $\text{Mg}(\text{TiO}_2)_2$ then the the number density for magnesium, titanium, and oxygen, related only to that material, will be predictors.

It is probable that such a direct measurement of a geometrical aspect would be a good predictor due to the amount of physical information. If RF were to be applied on to the entirety of the CIF file, it is probable that it would be a bad fit, due to the Bias-Variance-trade-off as mentioned in (REF), and because of the complexity of some of these files.

4.3 Void Fraction

Void Fraction, or the porosity, is a measurement of the void space in the material. Calculated *ab initio* with Poreblazer^{??}. We measure the accessible void, that is, the total amount of void space accessible from the surface. The pore volume is obtainable if Gurvich rule^{??} is fulfilled. It states that "if the density of the saturated nitrogen in the pores is assumed equal to its liquid density, regardless of the shape of the internal void network and, because of the weak interactions, regardless of the chemistry of the framework." The pore volume (v_{pore}) and the porosity (θ) are computed from:

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

$$\theta = v_{pore} \cdot \rho_{cryst} \quad (4.3)$$

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.

Two different pore volumes are calculated, the geometric pore volume, Ge_{pv} , which is defined as all the free volume of the unit cell, and Helium pore volume, He_{pv} , where the unit cell that can fit a probe with realistic intermolecular potential is tested. The calculation are done on the fixed 0K temperature.

Void Fraction is a good 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)

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:

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

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, I 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 manner, 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 discharged-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

MSE, PCA, R2, compair. Ordered by target:

Write a paragraph on how the "extra" entries affected the prediction result, and how you removed the once with less then 3-5 entries.

5.1 Random factors from database.

5.1.1 Average Voltage

5.1.2 Capacity

5.1.3 Energy Density

5.2 Volumetric number density

Volumetric number density or n as described(section ref), are shown(Tablature ref), for the Li- and Mg- databases. There are a couple of different results that are particularly interesting. First of all; n is overall better at predicting VC and with that GC naturally follows. In the first figure(??), the Mg-database, every prediction has a lower 'score' than in the Li-database. This is probably due to it being a much smaller database.

As can be seen clearly; the evaluations of this method is somewhat splitt. The MSE is generally better for the Li-database, and is best at the AV. The CV is better for the capacities but there is somewhat a high degree of uncertainty. This is probably due to the database being small. As expected the results for volumetric capacity is the peak of these runs.

One can from these results conclude that volumetric number density is worth bringing on as a predictor, as it clearly ha a part of the pusle.

<i>Target:→ Accuracy:↓</i>	AV	GC	VC	SE	ED
R^2 -score	0.5911	0.6815	0.6978	0.7103	0.5596
R^2 -train	0.9456	0.9279	0.9502	0.9241	0.9401
CV:	0.3180 (+/- 0.4977)	0.5423 (+/- 0.4435)	0.5943 (+/- 0.5877)	0.4494 (+/- 0.4608)	0.4755 (+/- 0.4464)
MSE:	0.4407	2491	41608	5284.5403	796956
CV-mean:	0.3076	0.5982	0.6511	0.4150	0.4700

Figure 8: Mg- prediction accuracy results 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(REF).

<i>Target:→ Accuracy:↓</i>	AV	GC	VC	SE	ED
R^2 -score	0.3309	0.3867	0.5631	0.3796	0.4027
R^2 -train	0.9193	0.9261	0.9286	0.9179	0.9266
CV:	-10.0739 (+/- 7.0309))	0.5009 (+/- 0.1844)	0.6095 (+/- 0.2123)	0.0341 (+/- 1.6924)	0.0298 (+/- 2.0854)
MSE:	0.7683	3318	29971	45911	532777
CV-mean:	-10.4347	0.5200	0.6465	0.4107	0.5093

Figure 9: Li- prediction accuracy results 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(REF).

5.2.1 Average Voltage

Mg db Charged:

Discharged: Mg db:

Li db:

5.2.2 Capacity

5.2.3 Energy Density

5.3 Void fraction

<i>Target:→ Accuracy:↓</i>	AV	GC	VC	SE	ED
R^2 -score	-0.0461	0.3426	0.3784	-0.0011	0.0838
R^2 -train	0.8676	0.8799	0.9052	0.8764	0.8786
CV:	-0.7976 (+/- 1.3547)	0.1844 (+/- 0.2182)	0.2983 (+/- 0.3636)	-0.3742 (+/- 0.9904)	-0.1623 (+/- 0.7009)
MSE:	1.1457	4521	62785	92593	1613762
CV-mean:	-0.5899	0.1660	0.3166	-0.2962	-0.0398

Figure 10: Mg- prediction accuracy results 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(REF).

5.4 AP-RDF

PCA, R2, MSE.

<i>Target:→ Accuracy:↓</i>	AV	GC	VC	SE	ED
R^2 -score	-0.1031	0.0263	0.0159	-0.0857	-0.1346
R^2 -train	0.8529	0.8434	0.8530	0.8519	0.8471
CV:	-0.7544 (+/- 1.2666)	-0.0272 (+/- 0.1926)	0.0916 (+/- 0.1639)	-0.4965 (+/- 0.9204))	-0.2823 (+/- 0.6630)
MSE:	1.2888	6426	110679	100803	1775980
CV-mean:	-0.6110	-0.0109	0.0897	-0.4422	-0.1690

Figure 11: Mg- prediction accuracy results 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(REF). With the AP-RDF as predictor.

5.5 Stability

This is a novel work, the aim is therefor to explore different predictors, by figuring out the different weights of the predictors on different targets, and which predictors that does not favorable for our predictions.

The model is decent at predicting; Gravimetric and Volumetric Capacity(87%), Specific Energy(70%), and Energy density(68%), but has no capability of predicting stability as of now. There is a need for *ab initio* calculations for several of our predictors, they calculate something that we know most definitely is correlated to the target without any premature calculations. This is something we are trying to move away from. As of now, only using the density fraction, we can get somewhere between 40% to 60% accuracy with our model.

This did not improve when including the void fraction, our predictions actually got worse.

AP-RDF - Still no good results! **This is a struggle.**

5.6 Geometrical descriptors

These graphs all represent the accuracy of the predictions on the training data and on new data given to the machine, with only the number density as a predictor, and the Average voltage, Gravimetric capacity, Volumetric capacity, energy density, and physical stability for the discharged material, as targets (a-f)???. Most notably the predictions on the Average voltage, Gravimetric capacity, Volumetric capacity, energy density, and specific energy are all showing a decent amount of correlation, with around 60% accuracy. The physical stability for the discharged-, and the physical stability of the charged-materials show that there is no correlation between the number density and the physical stability. It is also shown that there is no correlation between the number density and the void fraction. Or any of the other properties for that matter.

Part V

Summary

6 Summary and future work

6.1 Batteries

6.2 future work

6.2.1 improving method

References

- [1] International Energy Agency. “Global EV Outlook 2018: Towards Cross-modal Electrification”. In: IEA. 2018.
- [2] 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.
- [3] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.
- [4] Claus Daniel and Jürgen O Besenhard. *Handbook of battery materials*. John Wiley & Sons, 2012.
- [5] 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, Petersburg, &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.
- [6] *Example: Decision tree* Stefan Kottwitz. <http://www.texample.net/tikz/examples/decision-tree/>. Accessed: 2020-01-23.
- [7] James E Gentle, Wolfgang Karl Härdle, and Yuichi Mori. *Handbook of computational statistics: concepts and methods*. Springer Science & Business Media, 2012.
- [8] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [9] 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.
- [10] Stephen Marsland. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC, 2014.
- [11] Tom M Mitchell. *Machine learning*. 1997.
- [12] John P Perdew. “Density functional theory and the band gap problem”. In: *International Journal of Quantum Chemistry* 28.S19 (1985), pp. 497–523.
- [13] 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.
- [14] Thomas B Reddy. *Linden’s handbook of batteries*. Vol. 4. Mcgraw-hill New York, 2011.

-
- [15] 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.
- [16] 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.
- [17] *What is underfitting and overfitting in machine learning and how to deal with it.* Anup Bhande. https://miro.medium.com/max/2250/1*_70Pgojau8hkiPUiHoGK_w.png. Accessed: 2020-01-23.
- [18] Wikipedia. *Wikipedia, The Free Encyclopedia*. [Online; accessed 25-February-2020]. 2004. URL: https://upload.wikimedia.org/wikipedia/commons/thumb/5/56/Pila_di_Volta.jpg/800px-Pila_di_Volta.jpg0%7D.
- [19] Wikipedia. *Wikipedia, The Free Encyclopedia*. [Online; accessed 25-February-2020]. 2004. URL: https://upload.wikimedia.org/wikipedia/commons/thumb/a/a5/Galvanic_cell_labeled.svg/600px-Galvanic_cell_labeled.svg.png%7D.