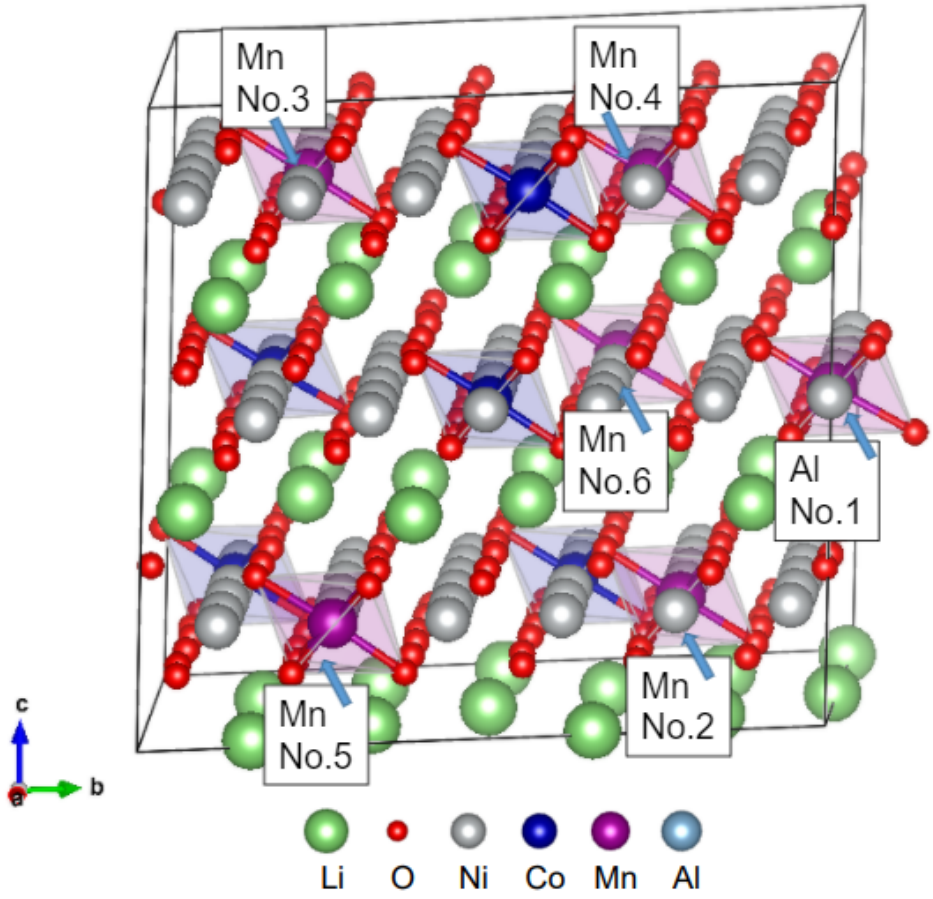
CHALMERS
UNIVERSITY OF TECHNOLOGY




# Predicting physical properties of NMC*M* cathode materials using machine learning guided DFT simulations

Master's thesis in Applied Physics

ALFRED STENSEKE

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021

Master's thesis 2021

# Predicting physical properties of NMC$M$ cathode materials using machine learning guided DFT simulations

ALFRED STENSEKE


CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2021

Predicting physical properties of NMC$M$ cathode materials using machine learning guided DFT simulations

ALFRED STENSEKE



Supervisors: Kazuki Higashi, Kunihiko Suzuki, Semiconductor Energy Laboratory Co., Ltd.
Examiner: Patrik Johansson, Department of Physics

Master's Thesis 2021
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Visualization of a NMCA molecule at 33% state of charge.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Predicting physical properties of NCM$M$ cathode materials using
machine learning guided DFT simulations
ALFRED STENSEKE
Department of Physics
Chalmers University of Technology

# Abstract

With the rapid increase in development of electric vehicles and energy storage systems, the demand for long lasting batteries with high energy density is higher than ever before. A crucial aspect of the market-leading lithium battery is the longterm cycling performance – to perform with high capacity even after thousands of charge-discharge cycles with as small degradation as possible. One cause for this degradation is the occurrence of small micro cracks in the cathode material due to small volume changes during charge-discharge cycles. To suppress this effect, state-of-the-art batteries today use metallic dopants such as aluminum in the cells of the cathode material. This project investigates other suitable dopants in NCM materials by implementing regression and gradient based prediction models on data acquired from supercomputer simulations using density functional theory (DFT). The results, while not fully conclusive, gives indications on what atomic features of dopants are interesting, as well as validates this relatively new machine learning approach in material science.

Keywords: DFT, NCM, lithium battery, machine learning, prediction.

# Acknowledgements

First and foremost I would like to thank everybody SEL for giving me the opportunity do this thesis with them – despite COVID trying its best to hinder it. I am especially grateful for all the insights and guidance provided by Kazuki Higashi and Kunihiko Suzuki, as well as the clear and responsive communication with Hinako Miyake. I would also like to thank Patrik Johansson for being my examiner, for mentoring me and for introducing me to the marvelous world of batteries.

Lastly, I would like to thank all my friends and family who have supported me not only during this project, but through my whole journey here at Chalmers – it would not have been possible without you.

Alfred Stenseke, Gothenburg, November 2021

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

Over the last decade there has been a high increase in development of electric vehicles (EVs) and large scale energy storage systems, and this trend is expected to continue (see figure 1.1) [1] [2]. EVs are perceived to be a more environmental friendly alternative to fossil fuel based vehicles and is part of the solution for many nation wide initiatives to lower carbon emissions. Central to the development and prosperity of EVs is the capability and sustainability of the battery. The ideal car battery would have a high energy density, rate capability, long cycle life and live up to various safety and geopolitical standards. Although achieving this has been the focus of many researchers in both academia and commercial industry, the complex nature of the atomic-scale interactions is not yet fully understood, and while contemporary batteries can thrive in several aspects they often lack in others. As of today, the lithium-ion battery (LIB) is the most used power source for such applications owing to its high energy density, longterm cycling performance and proficient rate capability [1].

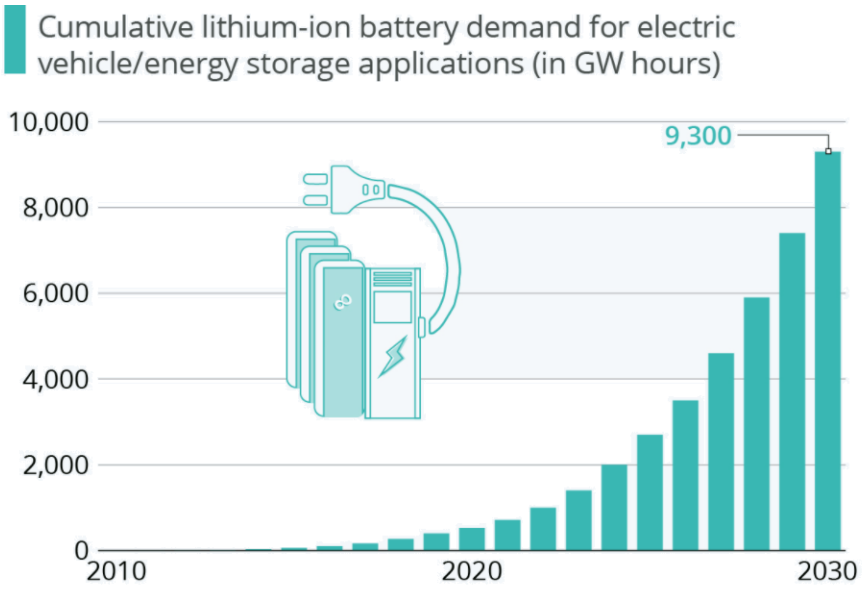



**Figure 1.1:** Expected trend for lithium-ion batteries in the coming decade.

## 1.1 Background

A lot of the focus in the development of new LIBs has been on the cathode material – where the Li-Ni-Mn-Co-$O_2$ (NMC or NCM) oxide, first developed by M. M Thackeray et al. in 2001, has provided some of the most successful results [3]. Before that, lithium cobalt oxide (LCO) cathodes dominated the market which caused a high demand on the troublesome cobalt metal. The introduction of Ni and Mn in the NMC battery cells shifted the cobalt dependency to a less troublesome nickel dependency [4]. At first, the NMC cells had equal parts of the three metals (NMC 111) but as research progressed, the amount of Ni increased to NMC 622 and even 811 [5]. The increase of Ni improves the capacity of NMC cells, but the method is limited by the decrease in long-term cycle performance and thermal stability for higher Ni concentrations [6]. One approach to overcome this is to dope the cathode with metallic atoms, with the most used dopant being Al – resulting in the state-of-the-art NMCA battery [7] [8].

There is a large amount of other potential dopants, $M$, for a NMC$M$ cathode material, especially when considering varying the concentrations of the different elements. It is not plausible to experimentally investigate all the various dopant configurations due to time and cost. However, using first principle density functional theory (DFT) simulations makes the search for promising configurations much more feasible. Such simulations are computationally expensive and would benefit greatly from machine learning algorithms aimed at effectively navigating the search space. [9]
Furthermore, while trivial tasks has been more and more automatized for a long time, the concept of automatizing scientific research is still rather new. To what extent can computer algorithms help scientists, not only when it comes to pure computations, but also extracting and understanding information? Going in to this project, the hypothesis is that machine learning can help reduce the amount of simulations/calculations required to investigate certain aspects of a material, and in doing so further increase the understanding of such materials.

## 1.2 Purpose

The purpose of this thesis is to propose battery material compounds with specific properties using machine learning driven simulations. The produced results should have predictive properties which can help increase the understanding of the underlying theory.

## 1.3 Aim

The aim of this thesis is to find a suitable dopant, $M$, for the lithium oxide cathode compound NMC$M$ to make the change in c-axis length as small as possible during the charge-discharge cycle. A machine learning algorithm will be trained in order to efficiently search the space of viable compound configurations.

## 1.4 Limitations

This thesis work will focus on finding a positive electrode material with a small volume change during charge/discharge cycles. Of course, there are many more aspects of what makes a proficient battery material, which will not be investigated. The project will be purely theoretical/computational and so the results will not be experimentally verified.

# 2
# Background & Methodology

The theory of the project can be divided into three different fields: lithium-ion batteries, computational methods with focus on DFT, and machine learning and neural network methods.

## 2.1 Lithium-ion Batteries

Lithium-ion (Li-ion) batteries exists in many different shapes, sizes and molecular configurations. They are widely used due to their high energy density and longterm cycle performance. The basic principle of all Li-on cell configurations is having two electrodes, the *anode* and the *cathode*, separated by an electrolyte. During discharge of the cell, Li-ions exits the cathode and flows toward the anode through the electrolyte. The freed electrons, previously associated with the Li-ions, travels toward the anode through the applied external circuit, where they perform the desired work. The Li-ions and the electrons are then intercalated into the anode. To charge the cell, an external voltage is applied which reverses the discharge process and the Li-ions returns to the cathode [10] [11].

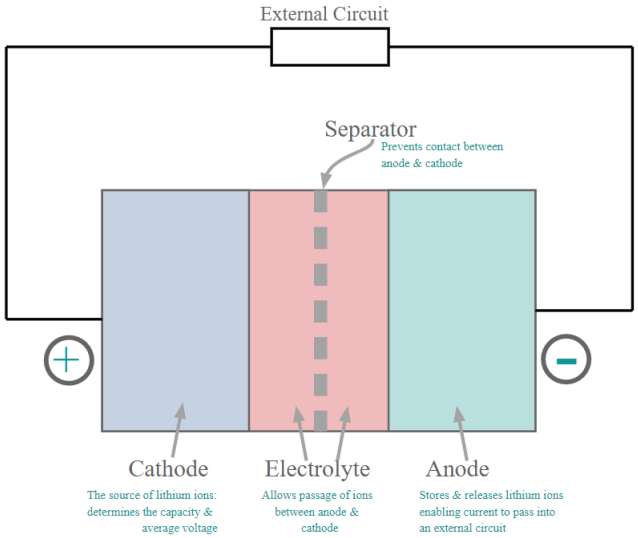



**Figure 2.1:** Simple overview of a battery.

While there is much theory and research regarding all parts of the battery, this project focuses exclusively on cathode materials. In the case of LIBs, the electricity is generated through the electrochemical reactions of lithium. Lithium, however, is not stable in the elemental form and is therefore combined with oxygen into lithium oxide in the cathode. The chemically active components, like lithium oxide for LIBs, are called the active material of the battery. A battery's voltage and capacity are dictated by the choice of active material: a larger difference in potential between anode and cathode yields a higher voltage, and a higher amount of lithium increases the capacity [12]. Present day LIB cells can deliver a voltage of 3.7 V and a battery energy density of 100-265 Wh/kg [10].

The layered rhombohedral structure (R3m) of $LiMO_2$ is what allows the repeated insertion and removal of ions to the cathode oxide [13]. During charge/discharge cycles the structure undergoes undesired distortions. Namely, the cell length along the c-axis changes a considerable amount. This change has been shown to cause degradation to the material in form of micro-cracks, which in turn lowers the performance of the cell over time [14] [15]. The c-axis change as a function of Li concentration can be seen in figure 2.2 for three different $LiMO_2$ materials.

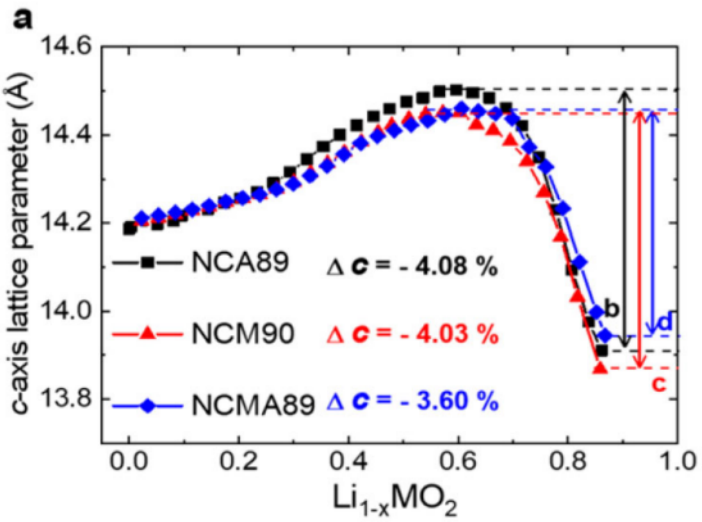



**Figure 2.2:** C-axis length as a function of lithium concentration in three different cathode materials.

The figure shows the difference between shortest and longest c-axis length of these three materials to be approximately 4%, where the longest point being close to x = 0.6 before rapidly dropping to the shortest point. The reason for the rapid drop starting at $x \approx 0.75$ originates from the opposing electrostatic repulsion and Van der Waals forces; for a lithiated state the electrostatic repulsion dominates whereas around the breakpoint of $x \approx 0.75$, the Van der Waals forces take over rapidly. Doping the metal oxide with a quaternary atom has shown to repress the c-axis change, as seen in the blue NCMA curve in figure 2.2 [14] [15].

## 2.2 Density functional theory

In theory, the many-body Schrödinger equation as seen in equation 2.1 contains all the information of a given quantum mechanical system:

$$\hat{H}\Psi = \left[\hat{T} + \hat{V} + \hat{U}\right]\Psi = \left[\sum_{i=1}^{N}\left(-\frac{\hbar^2}{2m_i}\nabla_i^2\right) + \sum_{i=1}^{N}V(\mathbf{r}_i) + \sum_{i<j}^{N}U\left(\mathbf{r}_i, \mathbf{r}_j\right)\right]\Psi = E\Psi, \tag{2.1}$$

where $N$ is the number of electrons, $\hat{T}$ is the kinetic energy, $\hat{V}$ is the potential energy and $\hat{U}$ is the interaction energy between electrons [16].

While it is feasible to solve the equation for very small systems, the computational power needed for larger systems is not realistically achievable. This is where DFT works as a method to approximate a solution for a many-body system. DFT methods are used to investigate electronic, magnetic and structural properties of molecules.

DFT reduces the dimensionality of the system through the Born-Oppenheimer approximation in which the nuclei of the molecules are seen as fixed compared to the high velocities of the electrons, greatly reducing the number of computations needed. Unlike methods like Hartree-Fock, DFT uses the density of the electrons as the fundamental property instead of dealing directly with the many-body wave function. This reduces the number positional variables in the wave function from three per atom in the system to a density function of only three variables. The Hohenburg-Kohn theorems asserts that all ground state properties can be determined using the electron density of the system. By using the electron density the system can be studied through a set of one-electron Schrödinger-like equations called Kohn-Sham equations. In addition, a so called Hubbard term, $U$, can be added to the total energy of the system to further improve the prediction of electron localisation [16] [17] [18].

## 2.3 Data analysis

Following the development of machine learning algorithms and its great success within other domains, data driven informatics strategies is now a promising tool in materials science. The approach enables predictions of a system based entirely on already acquired data rather than through experiments or simulations which are often costly and time-consuming. Indeed, the purely data-driven approach does not require any prior knowledge of the underlying system structure but instead uses statistical dependencies to distinguish interesting features.
Predictions normally involves extracting relevant *features* (also called *descriptors*) from existing input data and mapping the correlation between these and the property of interest. To validate the accuracy of the prediction, the data set is divided into train/test sets, normally at about 0.8/0.2 ratio. The mean squared error (MSE) between the predictions of the test data and the known output data is often used to quantify the accuracy of the prediction. The MSE is calculated as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{2.2}$$

where $n$ is the number of data points, $Y$ is the known target values of the test set and $\hat{Y}$ is the model predictions of the test set [19].
Several different prediction models where tested in this project and the following three yielded the most consistent results. They are all supervised learning models which means they learn mapping from inputs to already labeled outputs rather than extracting occurring patterns in input data, which is the case for unsupervised data.

### 2.3.1 Linear regression

Linear regression is one of the most basic prediction models. Given a set of data points with target values $Y$ (marked red in figure 2.3), it aims to find a regression line (blue) that minimizes the sum of squared residuals.

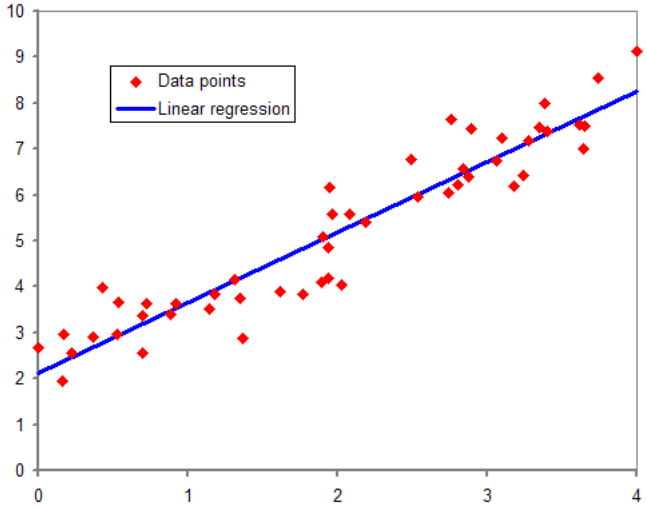



**Figure 2.3:** Basic illustration of a one-dimensional linear regression line. The model tries to find a line that minimizes the sum of squared distances between the line and the target values, marked as red.

The above example is of a one dimensional regression on the form $Y = a + bX$ but it works the same way for higher dimensions as follows

$$Y_i = a + b_1 X_{i1} + \cdots + b_p X_{ip}, \tag{2.3}$$

where $i$ denotes the data point and $p$ the number of included features.

Linear regression is straightforward, easy to understand and can be regularized to prevent overfitting. It does however struggle with more complex, non-linear relationships [20].

### 2.3.2 Random forest regression

Random forest regression is a ensemble learning method meaning it utilizes an ensemble of learning algorithms combined. It builds a "forest" of decision trees where the branch nodes are built up of subsets of parameter features. The ending nodes of the trees are called "leaves" and contains the target value of the individual paths. The average of many randomly picked decision trees is used to train the prediction model in order to circumvent overfitting and data variation problems.

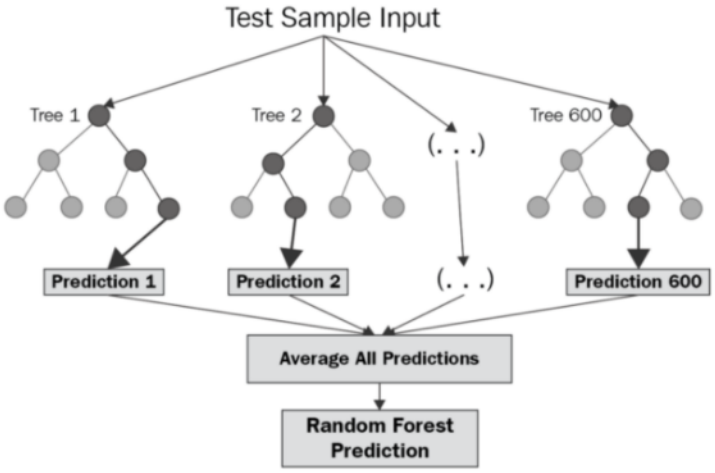



**Figure 2.4:** Overview of a random forest regression model. The average of many randomly picked decision trees makes up the final random forest prediction.

Random forest regression is a diverse model, applicable in many situations. It works well with both continuous and categorical data and does not require any normalization of data. It is however rather computationally expensive due to the combinations of many decision trees and offers little understanding of the importance of variables [21].

### 2.3.3 Gradient boosting (LightGBM)

In gradient boosting the idea is to iteratively add a small estimator to the prediction model to correct the MSE error. Given an initially weak model $F$ to be trained in $1 < m < M$ stages, the algorithm adds a new estimator $h_m$ such that $F_{m+1}(x) = F_m + h_m$ improves the estimate from the previous step. The algorithm will fit the $h$-term to the predecessing stage $y - F_m(x)$ residual.
Gradient boosting offers pros similar to random forest regression. It has the potential to improve accuracy even further than RFR but is more susceptible to noise and requires more hyperparameter tuning [22].

# 3
# Computational

The methodology of the project is divided into two main parts: the acquisition and the analysis of the data.

## 3.1 Acquisition of data

Initial DFT simulations were done using the software openMX on supercomputers provided by Semiconductor Energy Laboratory Co., Ltd. (SEL). During the these simulations, various aspects such as spin, Hubbard term and different initial potentials were investigated until adequately satisfactory results were reached compared to literature. After that, more extensive and thorough simulations were done using the software VASP. Eleven different doping elements were investigated with variations in dopant positions and concentrations. The target value of all simulations was the difference in c-axis length between the compositions.

### 3.1.1 Cell structure

The original cell is a NCMA89 model and is shown in figure 3.1. The cell served as a base for further simulations where the Al and up to two more Mn sites were substituted with the dopant of interest, resulting in lithium cathode molecule on the form $\mathrm{Li}_x\mathrm{Ni}_{0.89}\mathrm{Co}_{0.05}\mathrm{Mn}_y M_z$.

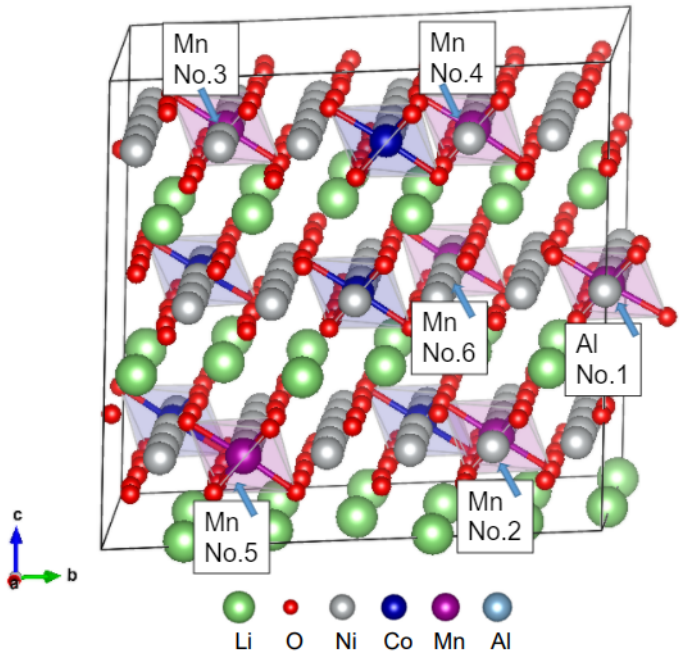



**Figure 3.1:** Overview of the baseline cell structure. Up to three of the six Mn/Al sites gets substituted by the dopant of interest during simulations. In this case, the cell is 33% lithiated.

All the positionally different permutations of a dopant were simulated to investigate the variations in c-axis length depending on dopant position. As an example: in the case of one Al atom, there are 6 different viable positions resulting in different c-axis lengths. For two Al atoms there are 15 different combinations of positions, and 20 different combinations for three Al atoms. Examples of the positional combinations can be seen in table 3.1

| | 1 dopant | 2 dopants | 3 dopants |
|---|---|---|---|
| Pos. Combs. | 1, 2, 3, 4, 5, 6 (6 total) | 1-2, 1-3, 1-4, 1-5, 1-6, 2-3, 2-4... (15 total) | 1-2-3, 1-2-4, 1-2-5, 1-2-6, 1-3-4... (20 total) |

**Table 3.1:** Table over the positional combinations of 1, 2 and 3 dopants.

### 3.1.2 Elements

The ten elements investigated in the project can be seen in figure 3.2 together with their place in the periodic table alongside the NCM atoms. As seen, the elements include most of the period 4 elements (fourth row on the periodic stable). These elements are all stable metals and many of them are very common in Earth's core/crust, making them suitable candidates for investigation [23]. While not period 4 elements, Mg and Al also fulfills these qualities.

| **Element:** | Mg | Al | Sc | Ti | Cr | Fe | Cu | Zn | Ga | Ge |
|---|---|---|---|---|---|---|---|---|---|---|
| **Atomic #:** | 12 | 13 | 21 | 22 | 24 | 26 | 29 | 30 | 31 | 32 |

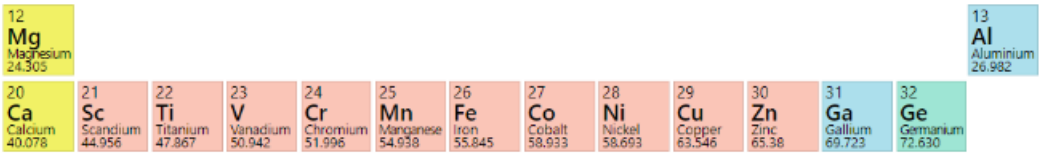



**Figure 3.2:** A table of the elements investigated as well as their place on the periodic table. Mn, Co and Ni are all included in NCM by default and are not investigated explicitly.

### 3.1.3 State of charge (SOC)

All of the above simulations could be computed for a multitude of states of charge. However, initial simulations as well as reference material indicated that the c-axis length is at a maximum around 33% lithium for all elements as shown for Al in figure 3.3. It also shows that the c-length minimum is at 0%. It would therefore give a good indication of the maximum $\Delta$c-length and save a lot of time/resources to only simulate at these two points. Because of deterioration implications of a fully discharged cell (high Ni batteries rarely goes under 20% SOC in real life applications), the minimum point was instead simulated at 16.7% Li (half of 33%) and $\Delta$c calculated as $c_{33\%Li} - c_{16.7\%Li}$.

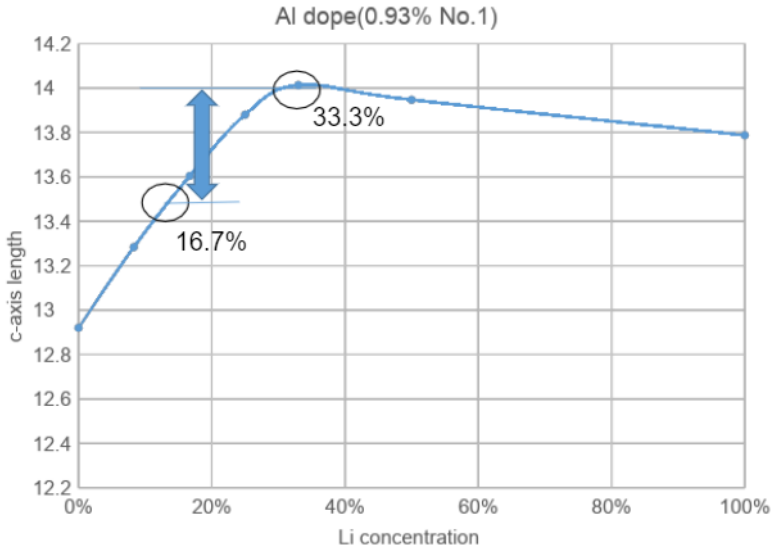



**Figure 3.3:** c-axis length as a function of SOC for a Al doped cell. $\Delta$c is calculated as the difference between 33% and 16.7% Li.

## 3.2 Data analysis

The data analysis was mainly done in Python using various open source libraries such as sklearn, seaborn, pandas and xenonpy.

### 3.2.1 Dataset

The simulations resulted in 410 different data points of $\Delta$c. An example of the structure of the data can be sin in figure 3.4a, where the 'No' column denotes the position of the dopants, 'delta_c' the target value, 'Type' the dopant element, and the rest are different features which may or may not be included after feature selection. The 'No' column was converted into a set of positional dummies, as seen in figure 3.4b, where 1/0 denotes if the position is occupied by a dopant or not.

| No | delta_c | Type | Valence | ionic_radius | M-O |
|---|---|---|---|---|---|
| 1 | -0.437 | Mg | 2 | 0.72 | 2.026 |
| 2, 3 | -0.44024989 | Zn | 2 | 0.74 | 2.0564 |
| 3, 5, 6 | -0.455342505 | Ga | 3 | 0.62 | 1.9922 |

**(a)** Three examples of simulated data points.

| | No1 | No2 | No3 | No4 | No5 | No6 | c |
|---|---|---|---|---|---|---|---|
| 0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.014744 |
| 1 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 14.025120 |
| 2 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 14.023130 |

**(b)** Example of the converted positional dummies. Note that the data points are not the same as in figure 3.4a.

The xenonpy library was used to extract the rest of the features. Xenonpy takes a molecular compound and returns 71 different element-level properties, resulting in an abundance of features [24].
Histograms of the data was plotted to visually investigate deviations in the target value distributions. Prediction models where trained both with and without the deviated data to explore its effect on the model accuracy.

### 3.2.2 Feature selection

The following cross validation approach was used to find the set of features resulting in the lowest prediction MSE while suppressing overfitting and susceptibility to data bias.
Starting with only the positional data seen in figure 3.4b as feature parameters, every parameter in the xenonpy library as well as the parameters in 3.4a was tried one by one. For every parameter, the data set was split into a train/test set with a 0.8/0.2 ratio using 20 different random seeds. The parameter resulting in the lowest average MSE of the 20 seeds was then added to the starting features and the process repeated itself until the MSE stopped improving by adding a feature. A simple overview of the approach can be seen in figure 3.5.

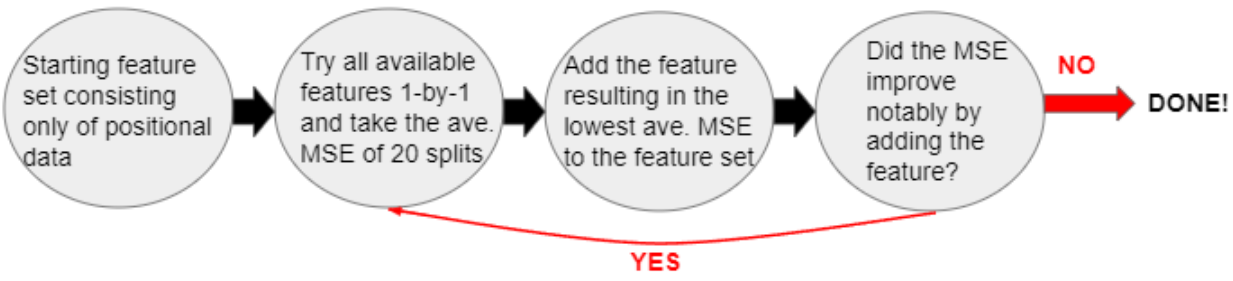



**Figure 3.5:** Flow chart of the feature selection. This approach suppresses overfitting and data bias.

This approach was done for three prediction models: linear Regression, random forest regression and lightGBM.

### 3.2.3 Prediction

In the last step, the three models were trained with features selected as above and used to predict $\Delta$c of an unknown data set. The unknown data set contained data points similar to the simulated data set, but with four or five doping atoms instead of up to three. The five unknown compounds with the lowest predicted $\Delta$c was then simulated to be compared to the models.

# 4

# Results

The simulated data returned a Gaussian distribution of the target values $\Delta$c with a mean value of -0.4598 Å and standard deviation 0.0197 Å. A non-negligible amount of incidents happened around -0.52 Å, as seen in figure 4.1a. Most of these came from simulations using Sc as dopant. Further data analysis was therefore made both including and excluding Sc. The distribution of the target values with all Sc simulations removed can be seen in figure 4.1b.

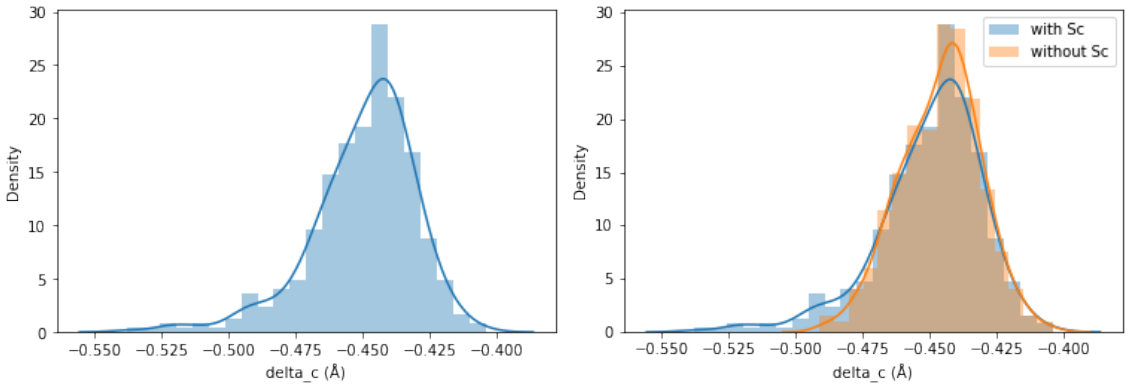



**Figure 4.1:** Distributions of the target values with and without Sc.

The data including Sc was made up of 410 data points and without Sc 370 data points.

## 4.1 Linear regression

The MSE resulting from the linear regression model after feature selection was 0.0002168 for the Sc data and 0.0001813 without Sc. The MSE progression after adding a new parameter can be seen in table 4.1. The MSE stopped improving after adding three parameters for the Sc data: *Atomic number*, *Melting point* and *polarizability*. For the data without Sc, the feature selection picked two different parameters than the Sc data: *heat capacity* and *van der waals radius*.

| Feature added | Data with Sc | Data without Sc |
|---|---|---|
| Only position | 0.0003387 | 0.0001949 |
| Atomic nmbr | 0.0002517 | - |
| Melting point | 0.0002247 | - |
| Polarizability | 0.0002183 | - |
| Heat capacity | - | 0.0001838 |
| vdw Radius | - | 0.0001813 |

**Table 4.1:** The improvement in MSE after adding the next best feature for the linear regression model.

A comparison of the distribution plots of predicted versus simulated values for an arbitrary train/test set can be seen in figure 4.2.

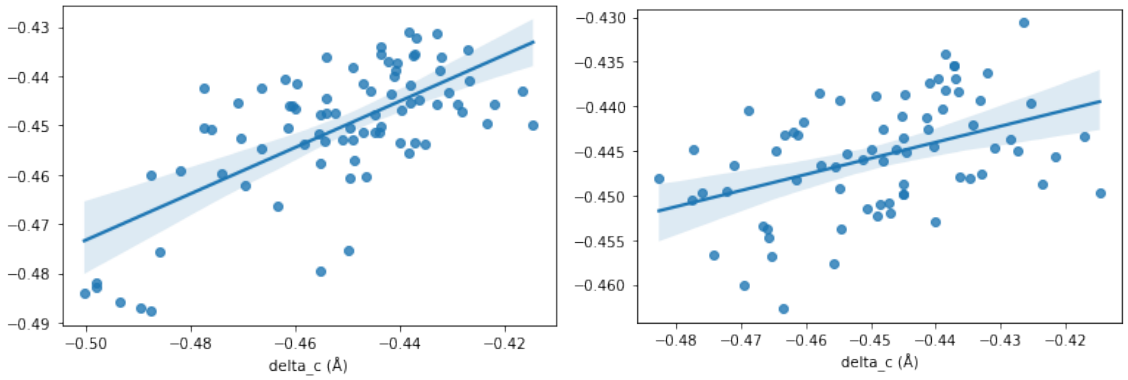



**Figure 4.2:** Predictions plotted against simulated values with (left) and without (right) the Sc values for the linear regression model. Note the difference in scale between the graphs.

The tendencies in the distributions are in large part the same, with most predictions between -0.46Å and -0.43Å. There is however a small but distinct cluster of under-predicted values around -0.48Å for the Sc data. This is further visualized in figure 4.3.

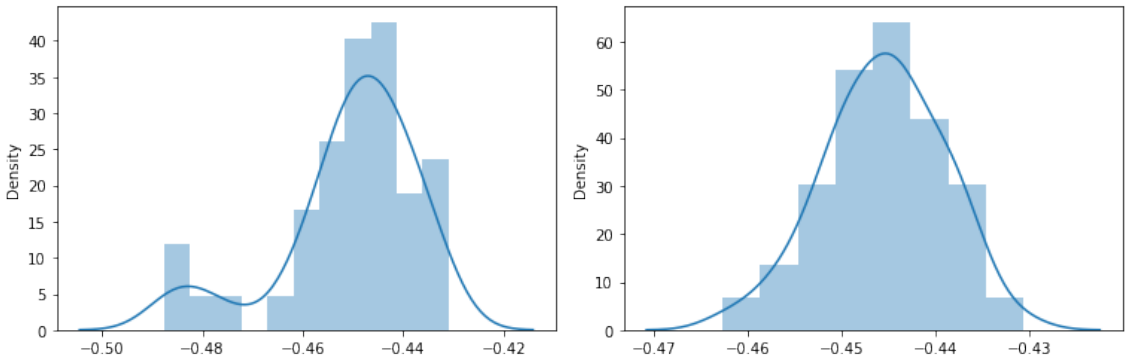



**Figure 4.3:** Distributions of the predicted target values with (left) and without (right) the Sc values using the linear regression model.

## 4.2 Random forest regression (RFR)

The MSE resulting from the random forest regression after feature selection was 0.0001520 for the data with Sc and 0.0001106 without Sc. The feature selection picked the same two features for both data sets: *Boiling point* and *number of electrons in d-shell*, with Sc data improving marginally from adding *DFT energy per atom (gs_energy)*.

| Feature added | Data with Sc | Data without Sc |
|---|---|---|
| Only position | 0.0003387 | 0.0001573 |
| Boiling point | 0.0001628 | 0.0001120 |
| # $e^-$ in d-shell | 0.0001547 | 0.0001106 |
| DFT energy/ atom | 0.0001520 | - |

**Table 4.2:** The improvement in MSE after adding the next best feature for the random forest regression model.

Figure 4.4 shows the prediction versus simulated values for the two data sets. While the Sc data for RFR does not have as many under-predicting points as for linear regression, some points under-predict to a larger extent. This trend is the same for all the train/test sets used and is further visualized when comparing the Sc data distributions between figures 4.3 and 4.5.

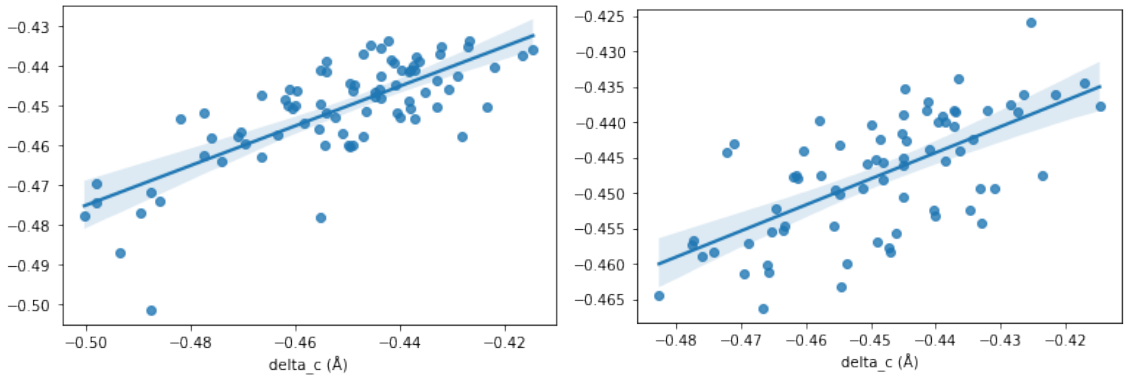



**Figure 4.4:** Predictions plotted against simulated values with (left) and without (right) the Sc values for the RFR model. Note the difference in scale between the graphs.

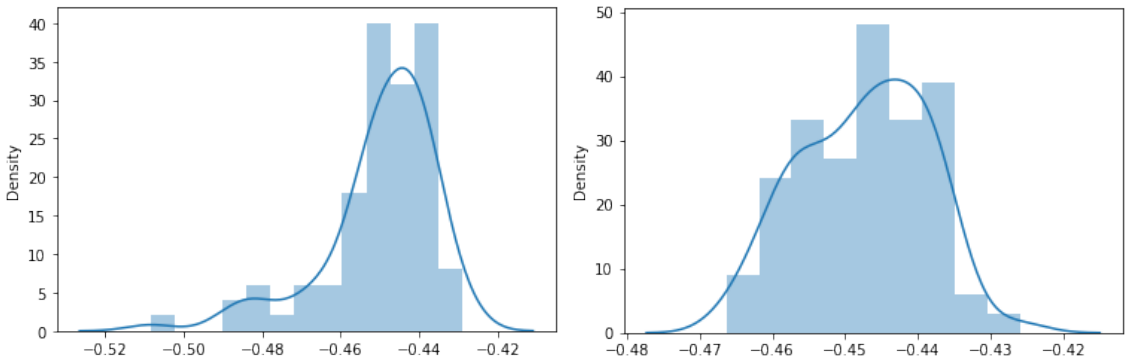



**Figure 4.5:** Distributions of the predicted target values with (left) and without (right) the Sc values using the RFR model.

## 4.3 LightGBM

The MSE resulting from the lightGBM model after feature selection was 0.0002484 for the data with Sc and 0.0001872 without Sc. The progression of the MSE with new parameters added can be seen in table 4.3. Both sets improved by adding *valence electrons* and the Sc set improved a bit further by adding *Heat capacity.*

| Feature added | Data with Sc | Data without Sc |
|---|---|---|
| Only position | 0.0003501 | 0.0001949 |
| Val. electrons | 0.0002523 | 0.0001872 |
| Heat capacity | 0.0002484 | - |

**Table 4.3:** The improvement in MSE after adding the next best feature for the lightGBM model.

As for previous methods, figure 4.6 shows the prediction versus simulated values for the two data sets, and figure 4.7 shows the distribution of the predicted target values $\Delta$c. Although requiring fewer features, the method does not seem to perform better than linear regression or RFR.

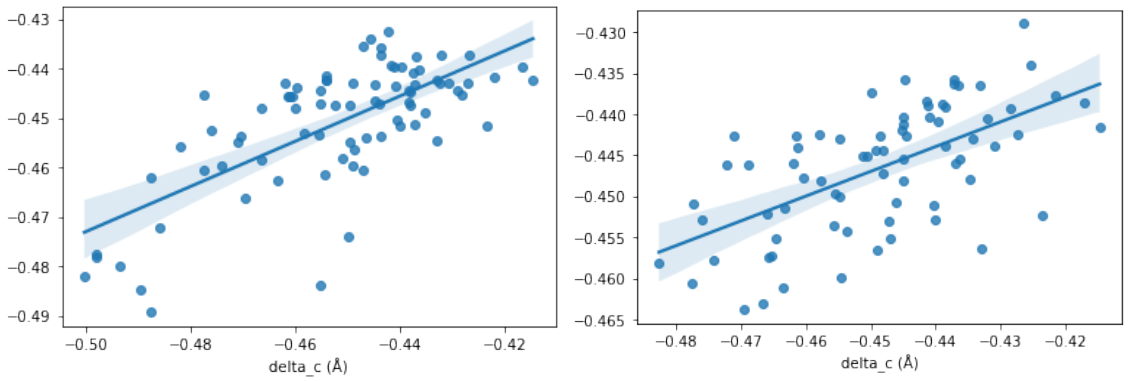



**Figure 4.6:** Predictions plotted against simulated values with (left) and without (right) the Sc values for the lightGBM model. Note the difference in scale between the graphs.

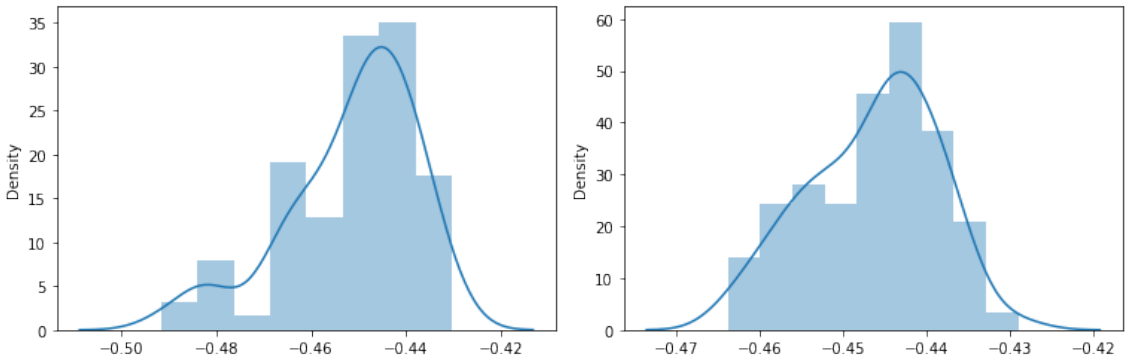



**Figure 4.7:** Distributions of the predicted target values with (left) and without (right) the Sc values using the lightGBM model.

## 4.4 Predictions

Predicting the compound with the lowest Δc from the unknown data set resulted in as seen in figure 4.8. When computed, the compounds returned values as seen in the last column.

| Linear-Regression | RandomForest-Regression | calculation |
|---|---|---|
| -0.44961719 | -0.43086003 | -0.4686577958 |
| -0.45091171 | -0.43092162 | -0.4434949766 |
| -0.45178979 | -0.43196576 | -0.4704425159 |
| -0.44624125 | -0.43381322 | -0.4371338258 |
| -0.45222402 | -0.43390011 | -0.4695949047 |

**Figure 4.8:** Predicted lowest Δc values from the unknown data set, as well as their later computed values.

# 5
# Discussion

Out of the three models, random forest regression performed the best with an average prediction error of about 2.3% on the data set without Sc. It also selected the same features for the sets with and without Sc which further validates the models success. Worth noting is that, while not a big issue for a data set of this size, RFR was considerably slower than the other models. Even though the RFR model had some success predicting test blocks of the simulated data set, it had little to no success predicting $\Delta$c of the unknown data set. Due to time limitations, no extensive investigation as to why the model performed poorly was made. There are however some observations to be made about the performance of the approach.

It is of course possible that the simulations are not consistent with reality both for the training set and the validations of the unknown set. Even if that is the case, they are hereon assumed to be correct for the sake of project progression.

## 5.1 Feature selection

At face value, the features appears somewhat arbitrary selected given how they vary not only between the models, but for the different data sets within the same model as well. A lot of the investigated features are, however, heavily correlated as indicated in figure 5.1. In the same sense *boiling point*, *heat capacity* or *melting point* are selected in all models and are all closely related. Furthermore, investigating 70 features for ten different elements probably resulted in a few "false" positive features. As an example, the Herfindahl-Hirschman Index (market competitiveness) improved the MSE substantially for a model.

| | |
|---|---|
| vdw_radius | Van der Waals radius |
| vdw_radius_alvarez | Van der Waals radius according to Alvarez |
| vdw_radius_batsanov | Van der Waals radius according to Batsanov |
| vdw_radius_bondi | Van der Waals radius according to Bondi |
| vdw_radius_dreiding | Van der Waals radius from the DREIDING FF |
| vdw_radius_mm3 | Van der Waals radius from the MM3 FF |
| vdw_radius_rt | Van der Waals radius according to Rowland and Taylor |
| vdw_radius_truhlar | Van der Waals radius according to Truhlar |
| vdw_radius_uff | Van der Waals radius from the UFF |

**Figure 5.1:** A section of the investigated features, all describing the Van der Waals radius in various ways.

Given more initial chemical knowledge, the number of investigated features could be heavily reduced; both the obviously unimportant and heavily correlated features could be removed. Even so, this purely data driven approach and its performance provides insight into the problem at hand.

## 5.2 Sc or no Sc

It is undeniably not ideal to exclude "bad" data points to improve the model accuracy. In this case, however, there was a clearly defined subset of data with outlying results and removing it improved the accuracy for all the models substantially. One plausible reason for the large c-axis change in Sc doped molecules is the large ionic radius of Sc. This causes the Sc atoms to move to the boundaries between the Li and Ni sites at low SoC instead of staying at the Ni sites like the other dopants. While the ionic radius of Sc is not *much* larger than other dopants, it appears to reach a certain breakpoint where this phenomenon occurs. In such case, the prediction models does not handle such discontinuities very well.

## 5.3 Models

It is inherently difficult for these three models to predict the very edge values from data with such narrow standard deviation. They would probably find more success trying to predict values well within the search space rather than on the very edges or outside: the models were trained using no more than 3 dopants, whereas the unknown data set had 4-6 dopants.

## 5.4 Conclusion

Rather than presenting a concrete suggestion of a doped NCM$M$ molecule with low $\Delta$c as it set out to do, the project finds its place as a proof of concept for a data driven approach to do so. Several interesting aspects were encountered during the process and the average error of 2.3% for the RFR model definitely validates the potential of the approach.

## 5.5 Future work

There is much more analysis to be done of the acquired data – both in tuning the investigated prediction models and trying new models and tools. Beyond that, more dopants from different periods of the periodic table can be explored and analysed. Finally, experimental verification of the predicted materials would further validate the approach and its results.

# Bibliography


[1] MarketsandMarkets. Lithium-ion battery market with covid-19 impact analysis, by type (li-nmc, lfp, lco, lto, lmo, nca), capacity, voltage, industry (consumer electronics, automotive, power, industrial), region (north america, europe, apac row) - global forecast to 2030: https://www.marketsandmarkets.com/market-reports/lithium-ion-battery-market-49714593.html, 2021.

[2] Nathaniel Bullard. This is the dawning of the age of the battery: https://www.bloomberg.com/news/articles/2020-12-17/this-is-the-dawning-of-the-age-of-the-battery?srnd=green, 12 2020.

[3] David Moore. Argonne lab's breakthrough cathode technology powers electric vehicles of today: https://www.batteriesinternational.com/2015/04/28/michael-thackeray/. *U.S. Department of Energy*, 2 2011.

[4] Bridget McCrea. Working to reduce cobalt dependency in battery manufacturing: https://www.sourcetoday.com/industries/article/21152018/working-to-reduce-cobalt-dependency-in-battery-manufacturing, 1 2021.

[5] Xin Sun, Xiaoli Luo, Zhan Zhang, Fanran Meng, and Jianxin Yang. Life cycle assessment of lithium nickel cobalt manganese oxide (ncm) batteries for electric passenger vehicles. *Journal of Cleaner Production*, 273, 11 2020.

[6] Hyung-Joo Noh, Sungjune Youn, Chong Seung Yoon, and Yang-Kook Sun. Comparison of the structural and electrochemical properties of layered li[nixcoymnz]o2 (x = 1/3, 0.5, 0.6, 0.7, 0.8 and 0.85) cathode material for lithium-ion batteries. *Journal of Power Sources*, 233, 7 2013.

[7] Fu Zhou, Xuemei Zhao, Zhonghua Lu, Junwei Jiang, and J.R. Dahn. The effect of al substitution on the reactivity of delithiated lini1/3mn1/3co(1/3z)alzo2 with non-aqueous electrolyte. *Electrochemistry Communications*, 10, 8 2008.

[8] S.-W. Woo, S.-T. Myung, H. Bang, D.-W. Kim, and Y.-K. Sun. Improvement of electrochemical and thermal properties of li[ni0.8co0.1mn0.1]o2 positive electrode materials by multiple metal (al, mg) substitution. *Electrochimica Acta*, 54, 6 2009.

[9] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3, 12 2017.

[10] Clean energy institute. What is a lithium-ion battery and how does it work? https://www.cei.washington.edu/education/science-of-solar/battery-technology/.

[11] Li-ion lithium ion battery. https://www.electronics-notes.com/articles/electronic_components/battery-technology/li-ion-lithium-ion-technology.php.
[12] The four components of a li-ion battery. https://www.samsungsdi.com/column/technology/detail/55272.html?pageindex=1idx=55272brdcode=001listtype=listsearchkeyword=.
[13] Yufang Chen, Chunman Zheng, Zhongxue Chen, and Kai Xie. The significance of the stable rhombohedral structure in li-rich cathodes for lithium-ion batteries. *Ionics*, 23, 2 2017.
[14] Tomohiro Yoshida, Kenta Hongo, and Ryo Maezono. First-principles study of structural transitions in linio <sub>2</sub> and high-throughput screening for long life battery. *The Journal of Physical Chemistry C*, 123, 6 2019.
[15] Un-Hyuck Kim, Liang-Yin Kuo, Payam Kaghazchi, Chong S. Yoon, and Yang-Kook Sun. Quaternary layered ni-rich ncma cathode for lithium-ion batteries. *ACS Energy Letters*, 4, 2 2019.
[16] Paul Erhart. A very short introduction to density functional theory (dft): http://physics.gu.se/ tfkhj/lecture_viii_dft-3.pdf.
[17] Theoretical Physics Group at University of Exeter. Density functional theory for beginners: http://newton.ex.ac.uk/research/qsystems/people/coomer/dft_intro.html.
[18] Elliott H. Lieb. The hubbard model: Some rigorous results and open problems: https://arxiv.org/pdf/cond-mat/9311033.pdf, 2004.
[19] Andrew Murphy and Candace Moore. Mean squared error, 6 2019.
[20] Astrid Schneider, Gerhard Hommel, and Maria Blettner. Linear regression analysis. *Deutsches Aerzteblatt Online*, 11 2010.
[21] Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*, 20, 3 2020.
[22] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 2013.
[23] Period 4 elements: https://en.wikipedia.org/wiki/period_4_element.
[24] Xenonpy features: https://xenonpy.readthedocs.io/en/stable/features.html.

**DEPARTMENT OF PHYSICS**
**CHALMERS UNIVERSITY OF TECHNOLOGY**
Gothenburg, Sweden
www.chalmers.se


CHALMERS
UNIVERSITY OF TECHNOLOGY