# Predicting solid-state qubit material host

by

Oliver Lerstøl Hebnes

Thesis

for the degree of

Master of Science

Faculty of Mathematics and Natural Sciences
University of Oslo

March 24, 2021

# Contents

# Part I

# Methodology and implementation

# Chapter 1

# Material Science Databases

There are multiple different databases for material science discovery available for every day use, some of them completely open-source while others are commercial. This chapter will give a brief overview of databases available for computational material science, and will serve as a toolbox for how to request information and what kind of python packages exist to process that information.

## 1.1 Fundamentals of a database

A quick search online will reveal the tremendous escalation of effort for big-data driven material science the last few years, resulting in several databases that stores ab-initio calculation details and results. We will here distinguish between a *cloud service*, which is a place to store independent databases for research and commercial purposes, and a *database*, which is an organized collection of structured information. As an example, a cloud service can store several databases, but a database cannot host a cloud service.

To limit the quest of databases, we have restricted the search for databases and cloud services to include inorganic compounds obtained experimentally or by first-principles calculations, in particular DFT-calculations using *Vienna ab initio simulation package* (VASP) [1]. VASP is a software for atomic scale materials programming. Table 1.2 and 1.3 shows a selection of databases and cloud services that meets the given criteries, respectively.

### 1.1.1 API and HTTP requests

To extract information from a database it is convenient to interact through an *API* (Application Programming Interface), which defines important variables such as the kind of requests to be made, how to make them and the data format for transmission. Importantly, this permits communication between different software medias. An API is entirely customizable, and can be made to extend existing functionality or tailormade for specific user-demanding modules.

The APIs that will be encountered is handled by the use of *HTTP* (Hypertext Transfer Protocol), which in its simplest form is a protocol that allows the fetching of resources. The protocol is client-server based, such as the client is requesting information and the server is responding to the request.

The most common HTTP-methods are GET, POST and HEAD, which are used to either retrieve, send, or get information about data, respectively. The latter request is usually done before a GET-method for requests considering large amount of data, since this can be a significant variable for the client's bandwith and load time. Following a request, the server normally responds with one of the status codes in table 1.1.

| Status code | Description |
|---|---|
| 2xx | OK - request was successful |
| 3xx | Resource was redirected |
| 4xx | Request failed due to either unsuccessful authentication or client error. |
| 5xx | Request failed due to server error. |

**Table 1.1:** Numeric status code for response. The leftmost digit decide the type of response, while the two follow-up digits depends on the implemented API.

A RESTful (Representational State Transfer) allows users to communicate with a server via a HTTP using a REST Architectural Style [2]. This enables the utilisation of Uniform Resource Identifiers (URI), where each object is represented as a unique resource and can be requested in a uniform manner. Importantly, this allows the use of both URIs and HTTP methods in an API, such that an object is represented by an unique URI whereas a HTTP-method can act on the object. This action will then return either the result of the action, or structured data that represents the object.

To provide a Python example, we can check the response by doing a GET request at the database Materials Project RESTful API in code listing 1.1. We use the preamble to version 2 of Materials Project, and add an API-check and an API-key. The response is shown in code listing 1.2. From the output, it is possible to tell that the supplied API-key is not valid, however, the request is valid.

```python
import requests
preamble = "https://www.materialsproject.org/rest/v2/"
url = preamble + "api_check"
```

```python
params = {"API_KEY":"unique_api_key"}
response = requests.get(url=url, params=params)
print(response.json())
```

**Listing 1.1:** Practical example of getting a response from Materials Project database.

```
{"valid_response": True,
"response":
  {"api_key_valid": False,
  "details":"API_KEY is not a valid key.",
  "version":
    {"db": "2020_09_08",
    "pymatgen": "2020.8.13",
    "rest": ""2.0"}
  }
}
```

**Listing 1.2:** Practical example of response from Materials Project request based on 1.1. The request was done 28. january 2020.

| Database | API | Free educational access | Number of entries |
|---|---|---|---|
| AFLOW | REST | True | 3.27 M |
| OQMD [3, 4] | RESTful API (qmpy, matminer) | True | 0.82 M |
| MP [5] | MAPI [6] | True | 0.71 M |
| ICSD [7] | RESTful API | False | 0.21 M |
| Jarvis-DFT | API | True | 0.04 M |

**Table 1.2:** A selection of databases of computational material science sorted after number of compounds. Abbreviations used are Novel Materials Discovery (NOMAD), Automatic-FLOW for Materials Discovery (AFLOW), Materials Project (MP), Inorganic Crystal Structure Database (ICSD) and Open Quantum Materials Database (OQMD). The number of entries can give the wrong perception of size of each respective database, as it does not visualise how many calculations have been done for each entry, nor if there might be duplicates.

### 1.1.2 Open-source Python libraries for material analysis

Many of the databases share convenient modules that are used to adapt, visualize, calculate or predict properties, making it easier for scientists to utilise

| Cloud service | API/REST | Open educational access |
|---|---|---|
| NoMaD | API | True |
| CMR [8] | ASE | RESTful API |
| MatNavi | API | True |
| PRISMS | REST | True |
| Citrine | API | True |
| MPDS | API | False |
| MDF | API | False |

**Table 1.3:** A selection of cloud services that offers database-storage. Abbreviations used are Computational Materials Repository (CMR), NIMS Materials Database (MatNavi), PRedictive Integrated Structural Materials Science (PRISMS), Materials Platform for Data Science (MPDS) and the Materials Data Fascility (MDF).

the databases. The Atomic Simulation Environment (ASE) is an environment in the Python programming language that includes several tools and modules for setting up, modifying and analyze atomistic simulations [9]. It is in particular used together with the cloud service Computational Materials Repository (CMR).

Another commonly used module is the Python Materials Genomics (pymatgen) [10]. This is a well-documented open module with both introductory and advanced use case examples written in Jupyter Notebook for easy reproducibility, and is integrated with the Materials Project RESTful API.

Another exceedingly popular library is matminer [11], which is an open-source toolkit for material analysis written in Python. Matminer is powered by a group known as *Hacking Materials Research Group* [1]. Matminer provides modules to extract data sets from many cloud-services and databases, with examples in table 1.2 and 1.3. Additionally, they provide the tools to extract possibly thousands of features from calculations based on DFT and more, and have modules for visualization and automatic machine learning. These tools will be examplified in the next chapter.

TODO : Add paragraph about sklearn.

A full selection of python libraries used and their versions can be found in the Github page (TODO: Add github page.)

[1] Project's Github site: https://github.com/hackingmaterials.

## 1.2 Databases and cloud services

Every database has its own speciality, and no two databases are the same. There exists entries that are fundamentally identical in several databases, but with different properties as a consequence of parameters used, such as the functional utilised in VASP or the relaxation scheme. This section digs up what exactly is each respective database's claim to fame.

### 1.2.1 Novel Materials Discovery

The Novel Materials Discovery (NOMAD) [12] Repository is an open-access platform for sharing and utilizing computational materials science data. NOMAD also consists of several branches such as NOMAD Archieve, which is the representation of the NOMAD repository parsified into a code-independent format, NOMAD Encyclopedia, which is a graphical user interface (GUI) for characterizing materials, and lastly NOMAD Analytics Toolkit, which includes early-development examples of artificial-intelligence tools [12].

Databases that are a part of NOMAD data collection includes Materials Project, the Open Quantum Materials Database and AFLOW. They are all based on the underlying quantum engine VASP.

### 1.2.2 Materials project

Materials project [5] is an open source project that offers a variety of properties of over one hundred thousand of inorganic crystalline materials. It is known as the initiator of materials genomics and has as its mission to accelerate the discovery of new technological materials, with an emphasis on batteries and electrodes, through advanced scientific computic and innovative design.

Every compound has an initial relaxation of cell and lattice parameters performed using a 1000k-point mesh to ensure that all properties calculated are representative of the idealized unit cell for each respective crystal structure. The functional GGA is used to calculate band structures, while for transition metals it is applied +U correction to correct for correlation effects in d- and f-orbital systems that are not addressed by GGA calculations [13]. The thermodynamic stability for each phase with respect to decomposition, is also calculated. This is denoted as E Above Hull, with a value of zero is defined as the most stable phase at a given composition, while larger positive values indicate increased instability.

Each material contains multiple computations for different purposes, resulting in different 'tasks'. The reason behind this is that each computation has a purpose, such as to calculate the band structure or energy. Therefore, it is possible to receive several tasks for one material which results in more features per material.

### 1.2.3 AFLOW

The AFLOW[14–16] repository is an automatic software framework for the calculations of a wide range of inorganic material properties. They utilise the GGA-PBE functional within VASP with projector-augmented wavefunction (PAW) potentials to relax twice and optimize the ICSD-sourced structur. They are using a $3000-6000$ k-point mesh, indicating a more computationally expensive calculation compared to the Materials Project. Next, the band structure is calculated with an even higher k-point density, in addition to the $+U$ correction term for most occupied d- and f-orbital systems, resulting in a standard band gap [17]. Furthermore, they apply a standard fit gathered from a study of DFT-computed versus experimentally measured band gap widths to the initial calculated value, obtaining a fitted band gap [18].

AFLOW-ML [19] is an API that uses machine learning to predict thermomechanical and electronic properties based on the chemical composition and atomic structure alone, which they denote as *fragment descriptors*. They start with applying a classification model to predict if a compound is either a metal or an insulator, where the latter is confirmed with an additional regression model to predict the band gap width. To be able to predict properties on an independent data set, they utilise a fivefold cross validation process for each model. They report a 93% prediction success rate of their initial binary classification model, whereas the majority of the wrongful predictions are narrow-gap semiconductors. The authors does not compare their predicted band gap to experimental values, but it is found that 93% of the machine-learning-derived values are within 25% of the DFT $+U$-calculated band gap width [20].

### 1.2.4 Open Quantum Materials Database

The Open Quantum Materials Database (OQDM) [3, 4] is a free and available database of DFT-calculations. It has included thermodynamic and structural properties of more than 600.000 materials, including all unique entries in the Inorganic Crystal Structure Database (ICSD) consisting of less than 34 atoms.

The DFT calculations are performed with the VASP software whereas the electron exchange and correlation are described with the GGA-PBE, while using the PAW potentials. They relax a structure using $4000-8000$ k-point mesh, indicating an even increasing computational expensive calculation than AFLOW again. Several element-specific settings are included such as using the $+U$ extension for various transition metals, lanthanides and actinides.

In addition, any calculation containing 3d or actinide elements are spin-polarized with a ferromagnetic alignment of spins to capture possible magnetism. However, the authors note that this approach does not capture complex magnetic, such as antiferromagnetism, which has been found to result in substantial errors for the formation energy [21].

### 1.2.5 JARVIS

Joint Automated Repository for Various Integrated Simulations (JARVIS) [22] - DFT is an open database based based on the VASP software to perform a variety of material property calculations. It consists of roughly 40.000 3D and 1.000 2D materials using the vdW-DF-OptB88 van der Waals functional, which was originally designed to improve the approximation of properties of two-dimensional van der Waals materials, but has also shown to be effective for bulk materials [23, 24]. The functional has shown accurate predictions for lattice-parameters and energetics for both vDW and non-vdW bonded materials [25].

Structures included in the data set are originally taken from the materials project, and then re-optimized using the OPT-functional. Finally, the combination of the OPT and modified Becke-Johnson (mBJ) functionals are used to obtain a representative band gap of each structure, since both have shown unprecedented accuracy in the calculation of band gap compared to any other DFT-based calculation methods [26].

The JARVIS-DFT database is part of a bigger platform that includes JARVIS-FF, which is the evaluation of classical forcefield with respect to DFT-data, and JARVIS-ML, which consists of 25 machine learning to predict properties of materials. In addition, JARVIS-DFT also includes a data set of 1D-nanowire and 0D-molecular materials, yet not publically distributed.

# Chapter 2

# Structural flow of information

The information stream of this project can be regarded as many modular parts connected together in logical pieces, and is strongly influenced by the process that defines a *minimum viable product* (MVP) through iterative development. An MVP is commonly known (in the bussiness world) as a new product that enables the most learning out of the minimum effort possible. This method allows a product to be iteratively evolved by consistent feedback and development, which in return enables cooperation between cross-disciplinary fields.

Furthermore, by having several modules serving as the fundament of the project, it is possible to achieve a long-lasting and robust product that is simple to maintain yet straightforward to develop. Bugs can be tackled through a documented code simultaneously as visible future improvements can be adressed. Therefore, the product is not regarded as completed in any terms, but rather ready for a first release after iteratively finding the mimimum viable product.

The main project of this work can be found on the Github repository *predicting-solid-state-qubit-candidates* [27], while the validation process can be adressed at the repository *predicting-ABO3-structures* [28]. In this chapter we will look into the details and thoughts behind the extraction of data, building features, data preparation, data mining and eventually fabricating a generalized model that can predict unseen data with confidence.

## 2.1 Extraction and featurization of data

The initial step for gathering and building features can be visualised through the flowchart in figure 2.1. Initially, we start by extracting all entries in the Materials Project that matches a specific query. Thereafter, we apply Matminer's featurization tools to make thousands of features of the data. In a parallel step, entries that are deemed similar to the entries from the initial Materials Project query are extracted from AFLOW, AFLOW-ML, JARVIS-

DFT, OQMD and Citrine Informatics. Finally, we combine the steps together as interim data and prepare the data for further analysis.

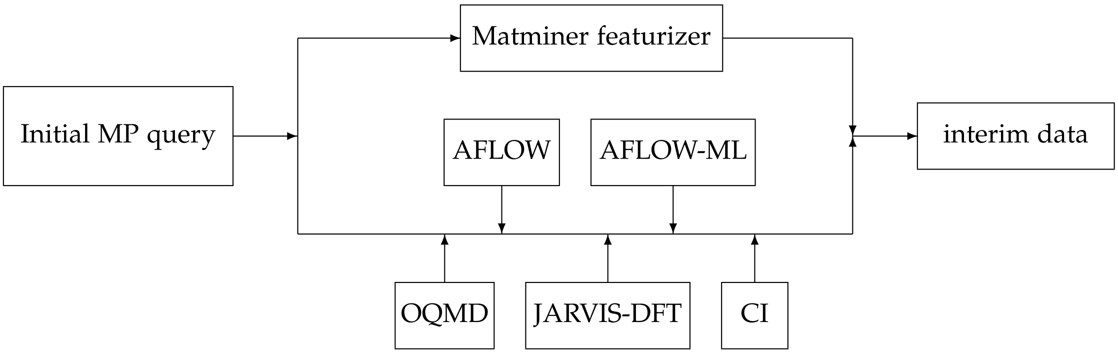

**Figure 2.1:** The data flow of the main project, starting from an initial MP-query, and ending with a featurized dataset with entries from several other databases. The matminer featurizer step is further visualized in-depth in figure 2.2.

The initial query has the requirement that all entries has to be derived from an experimental ICSD entry, and is reasoned by that we can identify equivalent entries in other databases. Furthermore, all entries in the Materials Project needs to have a band gap larger than 0.1eV. Recall that Materials Project applies the functional GGA in estimating the band gap, which is known to severely underestimate the given electronic property. Therefore, we have chosen a low value to not rule out any potential candidates but high enough to leave out all materials that can be considered metallic.

From figure 2.1 we notice that by using many databases we do not add additional entries that exist in some databases but is not to be found in Materials Project. This is by design since it preserves the versatility of choosing a database to work with. Therefore, one can completely ignore steps such as the initial query of Materials Project or the featurization process, and rather focus on e.g. all the 400.000 entries existing in OQMD. The examples that follows will illustrate the ease of extracting data from several different databases, and can serve as the starting point for other research projects in computational material science.

### 2.1.1 Practical data extraction with Python-examples

For this section, we will show practical examples of how to extract data that might fulfill the criteria for a material to host a qubit candidate given in the theory part. We will begin with the database of Materials Project, and then search for entries in other databases that match entries from MP. This process is reproducable as a jupyter notebook[1] and the databases in question are the ones refered to in the previous section.

Instead of building multiple HTTP-methods from scratch, we will here take a look at the easiest method at obtaining data from each database. This includes looking into the APIs that supports data-extraction and that are recommended by each respective database.

The range of data in a database can consist of data from a few entries up to an unlimited amount of entries with even further optional parameters, and has limitless use in applications. However, the amount of data in a database is irrelevant if the data is inaccessible. Therefore, we provide a toolbox in how to extract information in the easiest way possible.

Every data extraction class is based on an abstract parent class, which is listed in code listing 2.1. The advantages of using a base parent class are many, since it improves the readability during code reviews and reduce the main barrier for understanding the underlying structure of a project, while utilising reusable components. Yet, the main advantage of using a base parent class is the fact that it can effortlessly be extended for further implementations since it provides a code skeleton.

The structure of extraction is centered around using the data extraction tools, and not understanding them. Therefore, we only show how to use them here, while the code is found in the Appendix.

```python
import abc
import pandas as pd
from typing import Optional, Iterable, Tuple, Dict
import os
__all__ = ("data_base", )

class data_base(abc.ABC):
    # TODO: ADD COMMENTS.
    data_dir :          Optional[str] = None
    raw_data_path :     Optional[str] = None
    interim_data_path : Optional[str] = None

    df :       Optional[pd.DataFrame] = None

    def _does_file_exist(self)-> bool:
        if os.path.exists(self.raw_data_path):
            print("Data path {} detected. Reading now...".format(self.raw_data_path))
            return True
        else:
            print("Data for MP not detected. Applying query now...")
            return False #self.get_data()
```

[1] add and insert DOI for JN 01-generateDataset-notebook.ipynb

```

    def get_dataframe(self, sorted: Optional[bool] = True)->
pd.DataFrame:

        if self._does_file_exist():
            self.df = pd.read_pickle(self.raw_data_path)
        else:
            self.df = self._apply_query(sorted=sorted)
        print("Done")
        return(self.df)
```

**Listing 2.1:** Base parent class of all data extraction classes.

### Materials Project

The most up-to-date version of Materials Project can be extracted using the python package pymatgen, which is integrated with Materials Project REST API. Other retrievel tools that is dependent on pymatgen includes matminer, with the added functionality of returning a pandas dataframe. Copies of Materials Project are added frequently to cloud services such as Citrine Informatics, but the latest added entries to Materials Project cannot be guaranteed in such a query.

Entries in Materials Project are characterized using more than 60 features[2], some features being irrelevant for some materials while fundamental for others. The data is divided into three different branches, where the first can be described as basic properties of materials including over 30 features, while the second branch describes experimental thermochemical information. The last branch yields information about a particular calculation, in particular information that's relevant for running a DFT script.

To extract information from the database, we will be utilising the module pymatgen. This query supports MongoDB query and projection operators[3], resulting in an almost instant query.

1. Register for an account[4], and generate a secret API-key.
2. Set the required critera.
3. Set the wanted properties.
4. Apply the query.

[2] All features can be viewed in the documentation of the project: https://github.com/materialsproject/mapidoc/master/materials

[3] https://docs.mongodb.com/manual/reference/operator/query/

[4] https://materialsproject.org

The code nippet in code listing 2.2 resembles steps $2-4$, and is filtered as the inital query.

```
from src.data.get_data_MP import data_MP

MAPI_KEY = ‘‘very_secret_key_here’’
MP = data_MP(API_KEY=MAPI_KEY)
df = MP.get_dataframe()
```

**Listing 2.2:** Practical example of extracting information from Materials Project using pymatgen, resulting in a Pandas DataFrame named entries that contains the properties given after performing a filter on the database. The criteria is given as a JSON, and supports MongoDB operators.

### Citrine Informatics

Citrine Informatics is a cloud service, which means that the spectrum of stored information varies broadly. We will access research through open access for institutional and educational purposes. Information in Citrine can be stored using a scheme that is broken down into two sections, with private properties for each entry in addition to common fields that are the same for all entries.

In this example, we will gather experimental data using the module matminer. The following steps are required to extract information from Citrine Informatics.

1. Register for an account[5], and generate a secret API-key.
2. Set the required critera.
3. Set the wanted properties and common fields.
4. Apply the query.

The code listed in code listing 2.3 gives an easy example to steps $2-4$ with experimental data as filter.

```
from src.data.get_data_Citrine import data_Citrine

CAPI_KEY = ‘‘very_secret_key_here’’
citrine = data_Citrine(API_KEY=CAPI_KEY)
df = citrine.get_dataframe()
```

[5] https://citrination.com

**Listing 2.3:** Practical example of extracting information from Citrine Informatics using matminer, resulting in a Pandas DataFrame named experimental_entries that contains the properties given after performing a filter on the database. The criteria is given as a JSON.

### AFLOW

The query from AFLOW API [14] supports lazy formatting, which means that the query is just a search and does not return values but rather an object. This object is then used in the query when asking for values. For every object it is neccessary to request the desired property, consequently making the query process significantly more time-demanding than similar queries using APIs such as pymatgen or matminer for Citrine Informatics. Hence, the accessibility is strictly limited to either searching for single compounds or if the user possess sufficient time.

Matminer's data retrievel tool for AFLOW is currently an ongoing issue [29], thus we present in code listing 2.4 a function that extracts information from AFLOW and returns a Pandas DataFrame. In contrast to Materials Project and Citrine Informatics, AFLOW does not require an API-key for a query, which reduces the amount of steps to obtain data. The class searches for an stored AFLOW-data, and initialises a MP-query with the initial criteria if not successful. The resulting query will then be used as input to AFLOW.

```python
from src.data.get_data_AFLOW import data_AFLOW

AFLOW = data_AFLOW()
df = AFLOW.get_dataframe()
```

**Listing 2.4:** Practical example of extracting information from AFLOW. The function can extract all information in AFLOW for a given list of compounds, however, it is a slow method and requires consistent internet connection.

### AFLOW-ML

In this part, we will be using a machine learning algorithm named AFLOW-ML Property Labeled Material Fragments (PLMF) [19] to predict the band gap of structures. This algorithm is compatible with a POSCAR of a compound, which can be generated by the CIF (Crystallographic Information File) that describes a crystal's generic structure. It is possible to download a structure as a poscar by using Materials Project front-end API, but is a cumbersome process to do so individually if the task includes many structures. Extracting the feature of POSCAR is yet to be implemented in the RESful API of pymatgen, thus we demonstrate the versatility of pymatgen with a workaround.

We begin with extracting the desired compounds formula, its material_id for identification, and their respectful structure in CIF-format from Materials Project. In an iterative process, each CIF-structure is parsed to a pymatgen structure, where pymatgen can read and convert the structure to a POSCAR stored as a Python dictionary. Finally, we can use the POSCAR as input to AFLOW-ML, which will return the predicted band gap of the structure. This iterative process parsing and converting, but is an undemanding process. The function that handles this is presented in code listing 2.5. Similar to AFLOW-query, this code listing is dependent on MP-data and will apply for a query if the data is not present.

A significant portion of the process is tied up to obtaining the input-file for AFLOW-ML, and fewer structures will result in an easier process. Nevertheless, we present the following steps in order to receive data from AFLOW-ML.

1. Download AFLOWmlAPI[6].
2. Getting POSCAR from MP.
   (a) Apply the query from Materials Project with "CIF", "material_id" and "full_formula" as properties.
   (b) Insert resulting DataFrame into function defined in code listing 2.5.
3. Insert POSCAR to AFLOW-ML.

```python
from src.data.get_data_AFLOWML import data_AFLOWML

AFLOWML = data_AFLOWML()
df = AFLOWML.get_dataframe()
```

**Listing 2.5:** Practical example of extracting information from AFLOW-ML. The function will convert a CIF-file (from e.g. Materials Project) to a POSCAR, and will use it as input to AFLOW-ML. In return, one will get the structure's predicted band gap. It should be noted that this requires the AFLOW-ML library in the same directory.

### JARVIS-DFT

The newest version of the JARVIS-DFT dataset can be obtained by requesting an account at the official webpage, but with the drawback that an administrator has to either accept or deny the request. Thus, the accessibility of the

[6] http://aflow.org/src/aflow-ml/ to the same directory as code listing 2.5

database is dependent on if there is an active administrator paying attention to the requests, which is a limitation experienced during this work. Another approach is to download the database through matminer, however with the limitation of not neccessarily having the latest version of the database. A third approach is to download a version of JARVIS-DFT that have been made available for requests the 30.04.2020 at http://figshare.com by Choudhary *et al.* [22]. The author provides tools for extraction, yet not compatible with the latest version of Python (3.8) at the time writing (12.03.2021). Therefore, we provide a tool to extract this data through the use of our base class.

```python
from src.data.get_data_JARVIS import data_JARVIS

JARVIS = data_JARVIS()
df = JARVIS.get_dataframe()
```

**Listing 2.6:** Practical example of extracting information from JARVIS-DFT. For this example, we exclude all metals by removing all non-measured band gaps.

We observe that there is no advanced search filter when loading the database from matminer. The author of matminer regards this as the user's task, and is indeed easily done through the use of the python library Pandas.

## 2.2 Matminer featurization

Before applying any machine learning algorithm, raw data needs to be transformed into a numerical representation that reflects the relationship between the input and output data. This transformation is known as generating descriptors or features, however, we will in this work adapt the name *featurization*. The open source library of Matminer provides many tools to featurize existing features extracted from Materials Project. In this section we will describe how to extract the features from an initial Materials Project query result (see subsection. 2.1.1), and the resulting features. It is beyond the scope of this work to go in-depth of each feature since the resulting dataset contains a quantity of several thousand features, but we will here take the liberty to serve a brief overview of the features and refer to each respective citation for more information. The respective table with information regarding 39 distinct matminer featurizers is situated in the Appendix, table A.1.

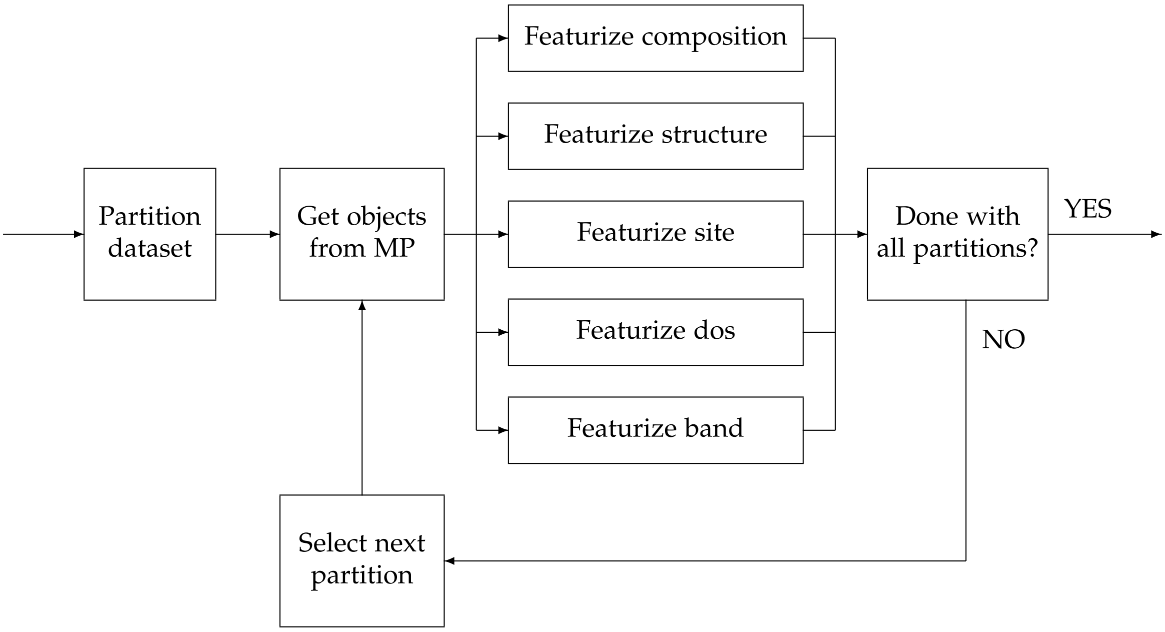

**Figure 2.2:** The process of the matminer featurizer step as seen in figure 2.1. To limit the memory and computational usage, the data is partioned into smaller subsets where the respective pymatgen objects are obtained through a query to be used in the following featurization steps. This is iteratively done until all the data has been featurized.

To apply matminer's featurization tools, we extend an existing implementation by De Breuck *et al.* [30] called the Materials Optimal Descriptor Network (MODNet). The author specifies that MODNet is a supervised machine learning framework for learning material properties based on either composition or crystal structure. To provide the training data for their model, MODNet featurizes (through matminer) structures either from Materials Project or in the form of a structure object made by pymatgen. Their current implementation provides featurization for compositions, structures and sites. However, matminer also provides featurization tools for density of states (DOS) and band structures, therefore we modify MODNet and extend it to fascilitate such featurizations.

One immediate limitation of our extension is that Matminer's tools is dependent on a pymatgen DOS- and bandstructure object. These objects contains information up to 5MB, and becomes a challenge when dealing with data containing several thousand such objects. This is solved by the required features for matminer's featurization for a subsample of the data, followed by a featurization process of the same subsample. When the feaurization is done, we store the new features and throw away the pymatgen features. This is done iteratively for the entire data set. Thus, a compromise between applying several queries and storing information has been done. The scheme can be visualised as the flow chart seen in figure 2.2.

In the extended version of the featurization process, we eliminate all columns that does not have any entries with physical meaning. This is beneficial for several reasons, such as to reduce memory allocated and to preprocess the data. If there are entries existing with both physical and non-physical for the same column, we replace the non-physical meanings with $-1$ for recognition in a later step. Additionally, we convert columns that are categorical or lacks a numerical representation into a categorical portrayal. Thus, we strive to limit the neccessary steps for further processing of data into a machine learning algorithm.

Even if the first version of Matminer was released in 2016, many issues concerning daily operational use are still present. During the featurization process in this work, we manually identified 14 (TODO: Update number) erroneous entries that are summarized in the Appendix, table A.2. These entries were excluded from the dataset.

# Chapter 3

# Data preparation and screening procedure

After the data has been selected by an initial query, followed by a thorough featurization process, we can finally investigate how the data looks like.

## 3.1

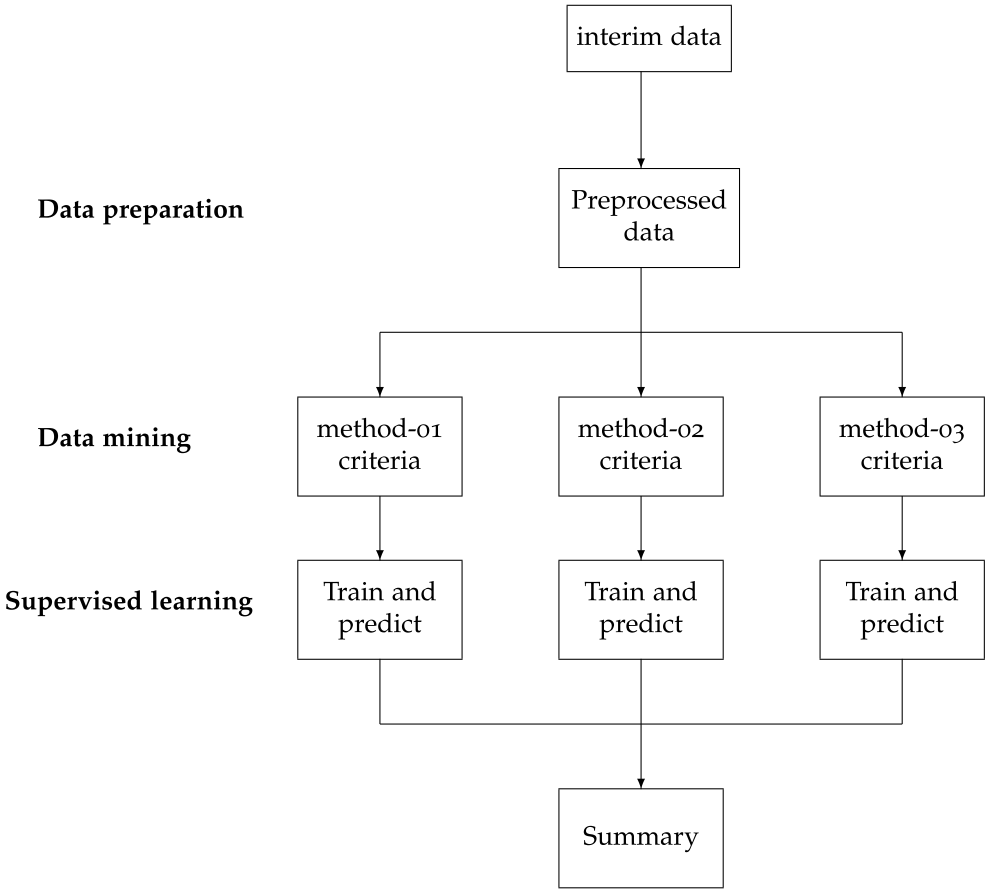
interim data
Data preparation
Preprocessed data
Data mining
method-01 criteria
method-02 criteria
method-03 criteria
Supervised learning
Train and predict
Train and predict
Train and predict
Summary

**Figure 3.1**

# Part II

# Appendices

# Appendix A

# Featurizaton

## A.1 Table of featurizers

**Table A.1:** This thesis' chosen 39 featurizers from matminer. Descriptions are either found from Ref. [11] or from the project's Github page.

| Features | Description | Original reference |
|---|---|---|

Continued on next page

**Table A.1 – continued from previous page**

| Features | Description | Original reference |
|---|---|---|
| **Composition features** | | |
| AtomicOrbitals | Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO). | [31] |
| AtomicPacking-Efficiency | Packing efficiency. | [32] |
| BandCenter | Estimation of absolute position of band center using geometric mean of electronegativity. | [33] |
| ElementFraction | Fraction of each element in a composition. | - |
| ElementProperty | Statistics of various element properties. | [10, 34, 35] |
| IonProperty | Maximum and average ionic character. | [34] |
| Miedema | Formation enthalpies of intermetallic compounds, solid solutions, and amorphous phases using semi-empirical Miedema model. | [36] |
| Stoichiometry | $L^p$ norm-based stoichiometric attributes. | [34] |
| TMetalFraction | Fraction of magnetic transition metals. | [35] |
| ValenceOrbital | Valence orbital attributes such as the mean number of electrons in each shell. | [34] |
| YangSolid-Solution | Mixing thermochemistry and size mismatch terms. | [37] |
| **Oxid composition features** | | |
| Electronegativity-Diff | Statistics on electronegativity difference between anions and cations. | [35] |
| OxidationStates | Statistics of oxidation states. | [35] |

Continued on next page

**Table A.1 – continued from previous page**

| Features | Description | Original reference |
|---|---|---|
| **Structure features** | | |
| DensityFeatures | Calculate density, volume per atom and packing fraction. | - |
| GlobalSymmetry-Features | Determines spacegroup number, crystal system (1-7) and inversion symmetry. | - |
| RadialDistribution-Function | Calculates the radial distribution function of a crystal system. | - |
| CoulombMatrix | Generate the Coulomb matrix, which is a representation of the nuclear coulombic interaction of the input structure. | [38] |
| PartialRadial-Distribution-Function | Compute the partial radial distribution function of a crystal structure | [39] |
| SineCoulomb-Matrix | Computes a variant of the coulomb matrix developed for periodic crystals. | [40] |
| EwaldEnergy | Computes the energy from Coulombic interactions based on charge states of each site. | [41] |
| BondFractions | Compute the fraction of each bond in a structure, based on nearest neighbours. | [42] |
| Structural-Heterogeneity | Calculates the variance in bond lengths and atomic volumes in a structure. | [43] |
| MaximumPacking-Efficiency | Calculates the maximum packing efficiency of a structure. | [43] |
| ChemicalOrdering | Computes how much the ordering of species differs from random in a structure. | [43] |
| XRDPowder-Pattern | 1D array representing normalized powder diffraction of a structure as calculated by pymatgen. | [10] |

Continued on next page

**Table A.1 – continued from previous page**

| Features | Description | Original reference |
|---|---|---|
| **Site features** | | |
| AGNI-Fingerprints | Calculates the product integral of RDF and Gaussian window function | [44] |
| AverageBond-Angle | Determines the average bond angle of a specific site with its nearest neighbors using pymatgens implementation. | [45] |
| AverageBond-Length | Determines the average bond length between one specific site and all its nearest neighbors using pymatgens implementation. | [45] |
| BondOrientational-Paramater | Calculates the averages of spherical harmonics of local neighbors | [46, 47] |
| ChemEnvSite Fingerprint | Calculates the resemblance of given sites to ideal environment using pymatgens ChemEnv package. | [48, 49] |
| Coordination-Number | The number of first nearest neighbors of a site | [49] |
| CrystalNN-Fingerprint | A local order parameter fingerprint for periodic crystals. | - |
| GaussianSymm-Func | Calculates the gaussian radial and angular symmetry functions originally suggested for fitting machine learning potentials. | [50, 51] |
| GeneralizedRadial-Distribution-Function | Computes the general radial distribution function for a site | [46] |
| LocalProperty-Difference | Computes the difference in elemental properties between a site and its neighboring sites. | [43, 45] |
| OPSite-Fingerprint | Computes the local structure order parameters from a site's neighbor environment. | [49] |

Continued on next page

**Table A.1 – continued from previous page**

| Features | Description | Original reference |
|---|---|---|
| Voronoi-Fingerprint | Calculates the Voronoi tessellation-based features around a target site. | [52, 53] |
| **Density of state features** | | |
| DOSFeaturizer | Computes top contributors to the density of states at the valence and conduction band edges. Thus includes chemical specie, orbital character, and orbital location information. | [54] |
| **Band structure features** | | |
| BandFeaturizer | Converts a complex electronic band structure into quantities such as band gap and the norm of k point coordinates at which the conduction band minimum and valence band maximum occur. | - |

## A.2 Erroneous entries

| MPID | Full formula | Reference |
|---|---|---|
| mp-555563 | $PH_6C_2S_2NCl_2O_4$ | [55] |
| mp-583476 | $Nb_7S_2I_{19}$ | [56] |
| mp-600205 | $H_{10}C_5SeS_2N_3Cl$ | - |
| mp-600217 | $H_{80}C_{40}Se_8S_{16}Br_8N_{24}$ | - |
| mp-1195290 | $Ga_3Si_5P_{10}H_{36}C_{12}N_4Cl_{11}$ | - |
| mp-1196358 | $P_4H_{120}Pt_8C_{40}I_8N_4Cl_8$ | - |
| mp-1196439 | $Sn_8P_4H_{128}C_{44}N_{12}Cl_8O_4$ | - |
| mp-1198652 | $Te_4H_{72}C_{36}S_{24}N_{12}Cl_4$ | - |
| mp-1198926 | $Re_8H_{96}C_{24}S_{24}N_{48}Cl_{48}$ | - |
| mp-1199490 | $Mn_4H_{64}C_{16}S_{16}N_{32}Cl_8$ | - |
| mp-1199686 | $Mo_4P_{16}H_{152}C_{52}N_{16}Cl_{16}$ | - |
| mp-1203403 | $C_{121}S_2Cl_{20}$ | - |
| mp-1204279 | $Si_{16}Te_8H_{176}Pd_8C_{64}Cl_{16}$ | - |
| mp-1204629 | $P_{16}H_{216}C_{80}N_{32}Cl_8$ | - |

# Bibliography

1. Kresse, G. & Furthmüller, J. Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set. *Physical Review B* **54,** 11169–11186 (Oct. 1996).
2. Battle, R. & Benson, E. Bridging the semantic Web and Web 2.0 with Representational State Transfer (REST). *Journal of Web Semantics* **6,** 61–69 (Feb. 2008).
3. Kirklin, S. *et al.* The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **1.** doi:10.1038/npjcompumats.2015.10 (Dec. 2015).
4. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65,** 1501–1509 (Sept. 2013).
5. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1,** 011002 (July 2013).
6. Ong, S. P. *et al.* The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science* **97,** 209–215 (Feb. 2015).
7. Levin, I. *NIST Inorganic Crystal Structure Database (ICSD)* en. 2020. doi:10.18434/M32147.
8. Landis, D. D. *et al.* The Computational Materials Repository. *Computing in Science & Engineering* **14,** 51–57 (Nov. 2012).
9. Larsen, A. H. *et al.* The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **29,** 273002 (June 2017).
10. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (Feb. 2013).

11. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152,** 60–69 (Sept. 2018).

12. Draxl, C. & Scheffler, M. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials* **2,** 036001 (May 2019).

13. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA+U framework. *Physical Review B* **73.** doi:10.1103/physrevb.73.195107 (May 2006).

14. Curtarolo, S. *et al.* AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **58,** 227–235 (June 2012).

15. Curtarolo, S. *et al.* AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science* **58,** 218–226 (June 2012).

16. Calderon, C. E. *et al.* The AFLOW standard for high-throughput materials science calculations. *Computational Materials Science* **108,** 233–238 (Oct. 2015).

17. Setyawan, W. & Curtarolo, S. High-throughput electronic band structure calculations: Challenges and tools. *Computational Materials Science* **49,** 299–312 (Aug. 2010).

18. Setyawan, W., Gaume, R. M., Lam, S., Feigelson, R. S. & Curtarolo, S. High-Throughput Combinatorial Database of Electronic Band Structures for Inorganic Scintillator Materials. *ACS Combinatorial Science* **13,** 382–390 (June 2011).

19. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications* **8.** doi:10.1038/ncomms15679 (June 2017).

20. Ferrenti, A. M., de Leon, N. P., Thompson, J. D. & Cava, R. J. Identifying candidate hosts for quantum defects via data mining. *npj Computational Materials* **6.** doi:10.1038/s41524-020-00391-7 (Aug. 2020).

21. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B* **85.** doi:10.1103/physrevb.85.115104 (Mar. 2012).

22. Choudhary, K. *et al.* JARVIS: An Integrated Infrastructure for Data-driven Materials Design. arXiv: `2007.01831v1 [cond-mat.mtrl-sci]` (July 3, 2020).

23. Thonhauser, T. *et al.* Van der Waals density functional: Self-consistent potential and the nature of the van der Waals bond. *Physical Review B* **76.** doi:10.1103/physrevb.76.125112 (Sept. 2007).

24. Klimeš, J., Bowler, D. R. & Michaelides, A. Van der Waals density functionals applied to solids. *Physical Review B* **83.** doi:10.1103/physrevb.83.195131 (May 2011).

25. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Physical Review B* **98.** doi:10.1103/physrevb.98.014107 (July 2018).

26. Choudhary, K. *et al.* Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Scientific Data* **5.** doi:10.1038/sdata.2018.82 (May 2018).

27. Ohebbi. *ohebbi/predicting-solid-state-qubit-candidates: v0.1-beta* 2021. doi:10.5281/ZENODO.4633959.

28. Ohebbi. *ohebbi/predicting-ABO3-structures: v0.1-alpha* 2021. doi:10.5281/ZENODO.4633968.

29. Rosenbrock, C. W. A Practical Python API for Querying AFLOWLIB. arXiv: 1710.00813v1 `[cs.DB]` (Sept. 28, 2017).

30. Breuck, P.-P. D., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. arXiv: 2102.02263v1 `[cond-mat.mtrl-sci]` (Feb. 3, 2021).

31. Kotochigova, S., Levine, Z. H., Shirley, E. L., Stiles, M. D. & Clark, C. W. Local-density-functional calculations of the energy of atoms. *Physical Review A* **55,** 191–199 (Jan. 1997).

32. Laws, K. J., Miracle, D. B. & Ferry, M. A predictive structural model for bulk metallic glasses. *Nature Communications* **6.** doi:10.1038/ncomms9123 (Sept. 2015).

33. Butler, M. A. & Ginley, D. S. Prediction of Flatband Potentials at Semiconductor-Electrolyte Interfaces from Atomic Electronegativities. *Journal of The Electrochemical Society* **125,** 228–232 (Feb. 1978).

34. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2.** doi:10.1038/npjcompumats.2016.28 (Aug. 2016).

35. Deml, A. M., O'Hayre, R., Wolverton, C. & Stevanović, V. Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression. *Physical Review B* **93.** doi:10.1103/physrevb.93.085142 (Feb. 2016).

36. Weeber, A. W. Application of the Miedema model to formation enthalpies and crystallisation temperatures of amorphous alloys. *Journal of Physics F: Metal Physics* **17,** 809–813 (Apr. 1987).

37. Yang, X. & Zhang, Y. Prediction of high-entropy stabilized solid-solution in multi-component alloys. *Materials Chemistry and Physics* **132,** 233–238 (Feb. 2012).

38. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **108.** doi:10.1103/physrevlett.108.058301 (Jan. 2012).

39. Schütt, K. T. *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **89.** doi:10.1103/physrevb.89.205118 (May 2014).

40. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **115,** 1094–1101 (Apr. 2015).

41. Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **369,** 253–287 (1921).

42. Hansen, K. *et al.* Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **6,** 2326–2331 (June 2015).

43. Ward, L. *et al.* Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B* **96.** doi:10.1103/physrevb.96.024104 (July 2017).

44. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *International Journal of Quantum Chemistry* **115,** 1074–1083 (Dec. 2014).

45. De Jong, M. *et al.* A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Scientific Reports* **6.** doi:10.1038/srep34256 (Oct. 2016).

46. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **95.** doi:10.1103/physrevb.95.144110 (Apr. 2017).

47. Steinhardt, P. J., Nelson, D. R. & Ronchetti, M. Bond-orientational order in liquids and glasses. *Physical Review B* **28,** 784–805 (July 1983).

48. Waroquiers, D. *et al.* Statistical Analysis of Coordination Environments in Oxides. *Chemistry of Materials* **29,** 8346–8360 (Sept. 2017).

49. Zimmermann, N. E. R., Horton, M. K., Jain, A. & Haranczyk, M. Assessing Local Structure Motifs Using Order Parameters for Motif Recognition, Interstitial Identification, and Diffusion Path Characterization. *Frontiers in Materials* **4.** doi:10.3389/fmats.2017.00034 (Nov. 2017).

50. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **134,** 074106 (Feb. 2011).

51. Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **207,** 310–324 (Oct. 2016).

52. Peng, H. L., Li, M. Z. & Wang, W. H. Structural Signature of Plastic Deformation in Metallic Glasses. *Physical Review Letters* **106.** doi:10.1103/physrevlett.106.135503 (Mar. 2011).

53. Wang, Q. & Jain, A. A transferable machine-learning framework linking interstice distribution and plastic heterogeneity in metallic glasses. *Nature Communications* **10.** doi:10.1038/s41467-019-13511-9 (Dec. 2019).

54. Dylla, M. T., Dunn, A., Anand, S., Jain, A. & Snyder, G. J. Machine Learning Chemical Guidelines for Engineering Electronic Structures in Half-Heusler Thermoelectric Materials. *Research* **2020,** 1–8 (Apr. 2020).

55. None Available. *Materials Data on PH6C2S2N(ClO2)2 by Materials Project* en. 2020. doi:10.17188/1268877.

56. None Available. *Materials Data on Nb7S2I19 by Materials Project* en. 2014. doi:10.17188/1277059.