Chapter 1

Experiments

In this chapter, the experiments which were carried out are described carefully, the results are reported and discussed. The first sections comments the computational method of the data retrieval, the second section describes the rocksalt-zincblende classification experiment and the last section describes the transparent semiconductor energy prediction experiment.

1.1 Density Functional Theory Data

In this section, the density functional theory data calculation procedure is described. Density Functional Theory (DFT) was studied in the previous work [citace bakalářky]. DFT is the way of tackling the many-body problem of quantum mechanics in exchange for the inevitable loss of calculation accuracy. This is done through the approximation of the many-body terms in the Hamiltonian of the Schrodinger's equation with a computationally advantageous approximation. The widely used Local Density Approximation (LDA) was used in the case of both our data sets [2], [3], [4], [5]. The authors of the data set used the software package FHI-aims [1] with tight settings in both cases [3], [4], [5] and periodic boundary conditions [4], [5]. The calculation procedure consists of initial guess of the atomic positions, iterative calculation of the electron density until desired precision is reached and relaxation of the atomic positions. The last two steps are repeated until desired accuracy is reached. During these steps, the target output properties are being calculated and in the end, reported by the software. The quantities can be further transformed to gain quantities with different physical meanings. It is now easy to see that the learning algorithms attempt to mimic this procedure and predict some of the output properties. Therefore, we do not predict the behavior of the real world quantum mechanical systems but their DFT approximations.

1.2 Rocksalt-Zincblende Classification Problem of Binary Compounds

The sources of this section are [2] and [3]. The main experiment carried out in the sources was replicated.

The goal of this experiment is to develop a model for prediction of the crystal structure of semiconductors. These compounds are simple: binary compounds AB consisting of element A and element B. The dataset consists of compounds which crystallize in three distinct structures: rocksalt (RS), zincblende (ZB) and wurtzite (WZ). However, the energies of ZB and WZ are very close and for the sake of simplicity, these two structures are not distinguished in this experiment. The target property of this experiment is the difference between the energy of rocksalt E_{RS} and the energy of zincblende E_{ZB} for

the given compound AB, that is $\Delta E_{AB} = E_{RS} - E_{ZB}$. Therefore, our goal is to find a model for binary compounds AB which assigns them the right crystal structure. The sign of the energy difference ΔE_{AB} gives the structure.

1.2.1 The Dataset

The dataset consists of 34 elements for which 7 physically meaningful features were calculated:

- Ionization potential IP [eV]
- Electron Affinity EA [eV]
- Highest Occupied Molecular Orbital H [eV]
- Lowest Unoccpuied Atomic Orbital L [eV]
- Radius where Radial Probability Density of valence s orbital is maximal r_s [Å]
- Radius where Radial Probability Density of valence p orbital is maximal r_p [Å]
- Radius where Radial Probability Density of valence d orbital is maximal r_d [Å]

In total 82 datapoints are available in this dataset (the table with the used data is available in Appendix A). It comes from quantum mechanics that these features are correlated in terms of the Pearson correlation coefficient of two 82-dimensional samples. The pairs (H, L) and (IP, EA) contain very similar information, namely the differences H - L and IP - EA but are not the same. The radii are as well correlated with the energy quantities. The correlation of the features is an issue and it will come up later on.

For the 82 compounds, the energy difference $\Delta E = E(RS) - E(ZB)$ was calculated. We put $\boldsymbol{y} = \boldsymbol{\Delta E}$ in accord with the notation in previous chapters. The labelling AB is not arbitrary. The label A is assigned to the element with lower Mulliken Electronegativity $EN = -\frac{1}{2}(EA + IP)$. This way, the so called primary features of a compound AB are defined as

$$\mathbf{x}_{AB} = \Big(IP(A), EA(A), H(A), L(A), r_s(A), r_p(A), r_d(A), IP(B), EA(B), H(B), L(B), r_s(B), r_p(B), r_d(B)\Big) \tag{1.1}$$

1.2.2 The Feature Space Generation

The nonlinear mapping $\Phi : \mathbb{R}^{14} \to \mathbb{R}^M$ is defined together with a set of unary and binary operations $\{|-|, +, /, \cdot, ()^2, \exp[]\}$ and the primary features $\mathbf{x}_{AB}$ are used to generate a feature space of expressions from which the matrix $\boldsymbol{X}$ is constructed. From this higher dimensial matrix, the optimal descriptors will be chosen. The mapping procedure introduces non-linearity needed for better description of the relationship between the primary features and the energy difference δE_{AB}

We will use the inequality (**??**) together with the knowledge that typically $C \in \langle 4, 8\rangle$ and this gives the estimate that a few thousand features is admissible. The primary features are divided into subsets based on the units of the features and meaning for more convenient illustration of the generation procedure.

ID	Quantities	Set size
A1	$IP(A), EA(A), IP(B), EA(B)$	4
A2	$H(A), L(A), H(B), L(B)$	4
A3	$r_s(A)$, $r_p(A)$, $r_d(A)$, $r_s(B)$, $r_p(B)$, $r_d(B)$	6

Table 1.1: The primary features divided into subsets based on their units and meaning

The features in (1.1) are then combined as follows and their number is calculated:

- B1: The sum and the absolute difference of two different features from A1
- B2: The sum and the absolute difference of different two features from A2
- B3: The sum and the absolute difference of two different features from A3
- C3: Squares of all A3 features and squares of all sums of features in B3
- D3: Exponentials of all A3 features and all sums in B3
- E3: Exponentials of C3
- F1: The following 4 expressions:

$$\begin{vmatrix} |IP(A) - EA(A)| + |IP(B) - EA(B)| \\ |IP(A) - EA(A)| - |IP(B) - EA(B)| \\ |IP(A) + EA(A)| + |IP(B) + EA(B)| \\ |IP(A) + EA(A)| - |IP(B) + EA(B)| \end{vmatrix}$$

- F2: The following 4 expressions:

$$\begin{vmatrix} |H(A) - L(A)| + |H(B) - L(B)| \\ |H(A) - L(A)| - |H(B) - L(B)| \\ |H(A) + L(A)| + |H(B) + L(B)| \\ |H(A) + L(A)| - |H(B) + L(B)| \end{vmatrix}$$

- F3: The same 4 expressions as in F1, F2 with inputs of all pairs of $r_s(A)$, $r_p(A)$, $r_d(A)$ in the first absolute term and all pairs of $r_s(B)$, $r_p(B)$, $r_d(B)$ in the second term.
- G: Ratios of all expressions in {Ai, Bi} with all expressions in {A3, C3, D3, E3} for i = 1, 2. The ratio 1/A3. Ratios A3/A3, A3/C3, B3/A3, B3/C3 such that only the unique expressions are chosen.

This gives total of 4376 potential descriptors and therefore 4376 columns of $\boldsymbol{X}$. The number of descriptors and examples for each set are given in table (1.2).

ID	Features	Set size
B1	$\lvert IP(A) + IP(B)\rvert$, $\lvert IP(B) - EA(B)\rvert$, ...	$2\binom{4}{2} = 12$
B2	$\lvert H(A) + H(B)\rvert$, $\lvert H(B) - L(B)\rvert$, ...	$2\binom{4}{2} = 12$
B3	$\lvert r_s(A) - r_p(A)\rvert$, $(r_d(B) + r_s(A))$, ...	$2\binom{6}{2} = 30$
C3	$r_s(A)^2$, $(r_d(B) + r_s(A))^2$, ...	$6 + \binom{6}{2} = 21$
D3	$\exp[r_s(A)]$, $\exp[r_d(B) + r_s(A)]$, ...	$6 + \binom{6}{2} = 21$
E3	$\exp[r_s(A)^2]$, $\exp[(r_d(B) + r_s(A))^2]$, ...	$6 + \binom{6}{2} = 21$
F1	$\Big\lvert \lvert IP(A) - EA(A)\rvert + \lvert IP(B) - EA(B)\rvert \Big\rvert$, ...	4
F2	$\Big\lvert \lvert H(A) - L(A)\rvert + \lvert H(B) - L(B)\rvert \Big\rvert$, ...	4
F3	$\Big\lvert \lvert r_s(A) - r_p(A)\rvert + \lvert r_s(B) - r_p(B)\rvert \Big\rvert$, ...	36
G	$\frac{IP(A)}{r_s(A)}, \frac{IP(A)}{(r_p(A)+r_s(B))^2}, \frac{IP(A)}{\exp[r_s(A)]}, \frac{\lvert r_p(A)-r_s(B)\rvert}{\exp[r_p(B)]}$, ...	4201

Table 1.2: Application of the chosen operations on primary features and the corresponding set sizes

1.2.3 The LASSO+ℓ_0 Method

As was briefly noted in the subsection (1.2.1), the correlation of the features is an issue. Additionally, we use combinations of already correlated quantities and therefore we can expect the correlation of the generated features to be a problem as well. Indeed, as it turns out LASSO itself performs unpredictably for values of λ which choose 1D ,2D ,3D and 4D descriptors and the optimal descriptor cannot be chosen this way reliably as we would do so for columns of a matrix with low correlation (see table (1.3)).

λ	Column index	#
0.0435	819	1
0.0410	819, 1105	2
0.0380	1105, 819, 1470	3
0.0375	1105, 819, 1470, 2172	4
0.0285	1105, 1470, 819, 2172, 966	5
0.0254	1105, 1470, 2172, 819, 966, 903	6
0.0252	1105, 1470, 2172, 819, 966, 1021, 903	7
0.0191	1105, 1021, 2540, 1470, 2172, 819, 903, 966	8
0.0122	2540, 2562, 3981, 1021. 2557, 1105, 2172, 253, 903	9
0.0121	2540, 2562, 3981, 1021, 2557, 1105, 2172, 253, 903, 819	10
⋮	⋮	⋮
0.0033	3399, 2562, 2769, 3110, 553, 3022, 1021, 2717, 3642, ...	15

Table 1.3: Found descriptors sorted from the most significant to the least for the λ value where their amount changes

The λ_i value is chosen using the following recursive formula

$$\lambda_i = \sqrt[dim(\lambda)-1]{\frac{1}{1000}}\lambda_{i-1}, i \in \{2, \ldots, dim(\lambda)\} \tag{1.2}$$

where $dim(\lambda)$ means the number of λ values evaluated during the LASSO+ℓ_0 procedure and $\lambda_1 = \frac{1}{N}\max_i |\langle \boldsymbol{x}_i, \boldsymbol{\Delta E}\rangle|$ is the threshold value when the first non-zero coefficient appears and where $\boldsymbol{x}_i$ is a column of $\boldsymbol{X}$. This formula was derived from the approaches to choosing λ values in [3].

The procedure appears to be reasonably stable for λ close the threshold λ_1. The significance of the first descriptor decreases with the descend of λ and it completely disappears for certain values of λ. The more non-zero coefficients are admissible, the less reliable the LASSO selection appears to be. For $\lambda \approx 0.003$, there are 15 non-zero coefficients and the selection is completely different compared to selections with higher λ values. Therefore, the following approach has been proposed [2]: The LASSO selection is carried out for a decreasing amount of λ values staring with λ_1. From these LASSO selections for various λ, the best Θ ocurring descriptors throughout the calculations are gather and among these the best 1D, 2D, 3D and 4D descriptor found using the least squares regression method for all $\binom{\Theta}{1}, \binom{\Theta}{2}, \binom{\Theta}{3}, \binom{\Theta}{4}$ which is effectively the ℓ_0 minimization. The 1D, 2D, 3D and 4D OLS model with the lowest MSEs are selected as the winners.

The issues and peculiarities of this approach applied to our dataset will be discussed in the following subsections.

1.2.4 Results and Discussion

The setting was chosen to be 100 values of λ starting from the threshold λ_1 and decreasing as a geometric sequence with $\lambda_{100} \approx 4.99 \cdot 10^{-5}$ being the lowest. The amount of contenders was capped at $\Theta = 30$. This will be further discussed later on. Before the training, the data were normalized using the built-in normalizer in the sci-kit learn LASSO implementation.

The LASSO+ℓ_0 method found the following models in the sense of minimal MSE:

$$\Delta E_{D1} = 0.055\frac{|IP(A) + IP(B)|}{r_p(A)^2} - 0.332$$

$$\Delta E_{D2} = 0.113\frac{|IP(B) - EA(B)|}{r_p(A)^2} - 1.558\frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} - 0.133$$

$$\Delta E_{D3} = 0.108\frac{|IP(B) - EA(B)|}{r_p(A)^2} - 1.807\frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} - 3.782\frac{|r_s(B) - r_p(B)|}{\exp(r_d(A))} - 0.023$$

$$\Delta E_{D4} = 0.188\frac{|H(A) + H(B)|}{\exp(r_p(A))^2} - 1.093\frac{|r_s(A) - r_p(B)|}{\exp(r_s(A))} - 4.680\frac{|r_s(B) - r_p(B)|}{\exp(r_d(A))} + 229.698\frac{|r_s(A) - r_d(B)|}{\exp(r_s(A) + r_d(B))} + 0.072$$

The starting point is the worst possible RMSE = 0.457 eV which is for prediction $\Delta E_{D_j} = 0$ for $j \in \{1, 2, 3, 4\}$ (the coefficients of the model are zero). RMSE and MaxAE of the models are reported in table (1.4).

Descriptor	1D	2D	3D	4D
RMSE [eV]	0.138	0.099	0.076	0.063
MaxAE [eV]	0.421	0.287	0.244	0.166

Table 1.4: Test RMSE and test MaxAE of the best k dimensional models

The 2D and 3D descriptors match exactly the ones from [2]. However, different coefficients were recovered. The 1D descriptor is does not match the one from [3] which is the first expression of 2D and 3D descriptor and our 1D descriptor is only slightly better. Interestingly, the 1D descriptor we found is the very first feature which LASSO recovers (labeled 819) whereas the other slightly worse feature (labeled

966) appears a couple steps later during the procedure (see table 1.3). The 4D descriptor was not reported in the publications [2] and [2]. However, we list it anyways. It is interesting to notice the second and the third terms in D4 match the second and the third terms in 3D. The test error RMSE and MaxAE drop most significantly between 1D and 2D descriptors. Also, the test MaxAE drops substantially between 3D and 4D descriptors.

The optimal choice of the amount of λ parameters and the amount of contenders Θ depends on the data set. In this case, there are 63 non-zero coefficients for the lowest value λ_{100}. Extending the geometric sequence up to λ_{150} yields up to 64 non-zero coefficients by the end of the procedure with very rapid growth of non-zero coefficients in the last 20 values of λ but the models did not change. The value of Θ was increased to 35, 40, 45, 50. The models recovered did not change after $\Theta = 35$. Examining Θ shows there are multiple entries of neighboring columns. Namely, it is 2557, 2558, 2559 and 3110, 3111, 3112. The Pearson correlation of pairs is between 0.83 and 0.92 for the first three neighboring descriptors and 0.91 and 0.997 for the second triplet. The features 3110 and 3111 have Pearson correlation of 0.997. Indeed, their form is very similar: $\frac{|r_s(B)-r_p(B)|}{\exp(r_d(A)+r_s(B))}$ and $\frac{|r_s(B)-r_p(B)|}{\exp(r_d(A)+r_p(B))}$.

1.2.5 Cross Validation, Sensitivity Analysis and Extrapolation

Given the small size of the data set, the cross validation approach of verification of the model was chosen. The data is split at random into two parts consisting of 10% and 90% of the data (this means 75 measurements for training and 7 measurements for testing). The model is learned on the train set and the RMSE is evaluated on the test set. This is repeated for 150 iterations and then the average RMSE over the 150 iterations is reported.

1.3 Transparent Conductors

<úvod o tom co je naše motivace>

1.3.1 Data Analysis