Magnesium Corrosion Data Exploration

Elisabeth Schiessler

February 12, 2021

1 The Dataset

The given dataset describes how certain additives influence magnesium corrosion of material *ZE41*. It contains 60 samples (additives/compounds). The first three columns of the dataset contain the compound name, the target value *inhibition efficiency ZE41 / %*, as well as a scaled version of the target column. Furthermore the dataset contains 1260 columns of features (molecular descriptors).

All data except for the compound name is given in decimal values, with varying ranges. As a preprocessing step, all descriptors are scaled to the interval $[0, 1]$ using linear (min-max) scaling per molecular descriptor. We directly use the scaled version of the inhibition efficiency, *LinIE ZE41*, as the target variable.

10% of the training data (i.e. 6 samples) are randomly selected and withheld from all analysis and training methods to be used as a validation set, c.f. table 1. This set is not varied at any point.

Table 1: Set aside validation set (randomly selected).

Index	Compound description
0	3-Amino-124-triazole
5	4-methylsalicylicacid
13	35-Dinitrosalicylicacid
36	maleicacid
45	para-toluicacid
54	salicylhydrocamicacid

2 Statistical Analysis

2.1 Searching for Individual Most Relevant Features

We aim to identify the $n \in \{1, 2, \ldots, 6\}$ most relevant molecular descriptors for predicting the target value. A popular means for feature selection is to train a Random Forest Regressor and inspect the features that are used as support, i.e. used for decision making in any of the calculated decision trees that make up the random forest. To counter deterministic influences, we train 100 times each on 100 different random seeds (i.e. 10.000 runs in total), each time permuting the order of the columns. The supporting features of all trained trees are stored, and those features that are chosen most often are deemed to be the most relevant ones. The result of this method can be seen in table 2.

To get some perspective on the relevance towards the target value of each individual feature selected by this method, the dataset is expanded by an additional column containing a random value, which is also reset for each calculation. The random column was included in the random forest's support in % of all runs.

Table 2: Most relevant features as determined by Random Forests, and how often they were included in the support.

Feature	Included in % of runs
VE2_G/D	20.26
Eig14_EA(dm)	20.25
Mor31m	20.10
TDB04u	20.03
HATS1e	19.99
random	19.30

2.2 Searching for Groups of Most Relevant Features

Instead of searching for individual most important features, we can also look for groups of n features that together most are able to most accurately predict, the target value. We do so by repeated application of recursive feature elimination (RFE). RFE searches for subsets of features by repeatedly fitting a chosen regression model, and discarding the least relevant features until only the desired number is retained. As the underlying regression model we chose a random forest regressor.

The RFE algorithm is run 100 times (with varying random seed to counter deterministic behaviour in any random seed generators). For each n the top 3 obtained lists of n most relevant molecular descriptors are then compared.

Table 3: Top 3 combinations per number of selected features as found by RFE. Third column indicates how often the given combination won the competition against potential other contestants.

n	Top 3 combinations per n	Won (x/100)
1	P_VSA_MR_5	71
	LUMO / eV	21
	Mor04m	8
2	P_VSA_MR_5, LUMO / eV	54
	P_VSA_MR_5, Mor04m	39
	LUMO / eV, Mor04m	5
3	P_VSA_MR_5, LUMO / eV, Mor04m	83
	P_VSA_MR_5, LUMO / eV, E1p	4
	P_VSA_MR_5, LUMO / eV, Mor22s	4
4	P_VSA_MR_5, LUMO / eV, Mor04m, E1p	29
	P_VSA_MR_5, LUMO / eV, Mor04m, Mor22s	27
	P_VSA_MR_5, LUMO / eV, Mor04m, P_VSA_LogP_2	13
5	P_VSA_MR_5, LUMO / eV, Mor04m, E1p, Mor22s	21
	P_VSA_MR_5, LUMO / eV, Mor04m, E1p, P_VSA_LogP_2	11
	P_VSA_MR_5, LUMO / eV, Mor04m, Mor22s, P_VSA_LogP_2	6
6	P_VSA_MR_5, LUMO / eV, Mor04m, E1p, P_VSA_LogP_2, Mor22s	11
	P_VSA_MR_5, LUMO / eV, Mor04m, E1p, Mor22s, HOMO / eV	8
	P_VSA_MR_5, LUMO / eV, Mor04m, E1p, P_VSA_LogP_2, HOMO / eV	6

As we can see from table 3, selecting the most relevant one to five molecular descriptors is a relatively clear decision, albeit with almost a tie for fourth place. Afterwards the number of times a combination wins the competition drops significantly.

When compared to the method used in section 2.1, out of the top 5 most relevant feature group, only

P_VSA_MR_5 individually has a higher relevance than the random variable, being used in 19.39 % of all supports.

2.3 Random Search

A third alternative is to randomly choose n features and train a deep learning model on it. Repeating this process often enough (e.g. 100 times for 100 different random seeds), and retaining the combinations of features that yields the best result will produce the best scoring model. However this method has to be viewed with some caution, as we can no longer check each trained model with regard to how much overfitting has occurred, and no fine-tuning of the produced models can happen due to sheer amount of repetitions. We thus perform this variant only for $n = 5$ to demonstrate its abilities, calling this variant MF, see section 3 for model specifics.

3 Deep Learning Models

To evaluate the results of the statistical analysis, we train several deep learning models and compare their performance on the validation set (c.f. table 1). We elect to train a total of 6 model variations:

- M3a - using the top 3 selected features from RFE as the input: P_VSA_MR_5, LUMO / eV, Mor04m
- M3b - using the top 3 individual most relevant features as the input: VE2_G/D, Eig14_EA(dm), Mor31m
- M5a - using the top 5 selected features from RFE as input: P_VSA_MR_5, LUMO / eV, Mor04m, E1p, Mor22s
- M5b - using the top 5 individual most relevant features as the input: VE2_G/D, Eig14_EA(dm), Mor31m, TDB04u, HATS1e
- MR - using the top 5 selected features from random search as the input: B04[C-C], CATS3D_03_DP, Eig11_AEA(ed), J_RG, SpMAD_EA(ed)
- MF - using the full dataset as the input

The models all consist of 3 hidden layers, with 50-20-10 units (M3, M5, MR) or 100-50-10 units (MF) in the hidden layers, using `relu` activation. They are trained for 25 epochs each, using an Adam optimizer with learning rates 0.01 (M3/M5/MR) and 0.005 (MF), and mean squared error as a loss function. These specifics were also used for the random search described in section 2.3. The obtained validation losses can be seen in table 4.

4 Results and Discussion

As expected, the individual most relevant features (M3b, M5b) do not necessarily work best as a group and thus yield very poor results. The 3- and 5-parameter models trained on the selected group features (M§a, M5a) on average perform better than models trained on 5 randomly selected features, as does the model trained on the full dataset (MF).

The best result is in fact obtained from randomly training a huge variation of deep learning models and then picking the best scoring one. However this approach is practical only due to the fact that the dataset is very small, containing only 60 samples. The dataset size should be kept in mind for any considerations regarding prediction quality. With only 60 samples, 10% of which were withheld from all training and evaluation processes, we can not expect any result to generalise well on unknown data.

Table 4: Achieved validation losses per deep learning model.

Model	Validation Loss
M3a	0.0190
M3b	0.0576
M5a	0.0174
M5b	0.0729
MR (avg)	0.0335
MR (best)	0.0052
MF	0.0185

Table 5: True vs. predicted values on the validation set as obtained by the deep learning models M3a, M5a and MF. All predicted values compared to *inhibition efficiency ZE41 / %* target.

Index	0	5	13	36	45	54
True value	-157.00	39.00	38.00	12.00	-6.00	-17.00
M3a	-97.85	13.73	132.81	23.88	-1.05	1.04
M3b	-18.07	-49.77	-60.47	-18.70	-49.60	-54.99
M5a	-107.34	8.83	122.16	54.16	-18.73	-12.51
M5b	44.34	-40.33	-16.44	-5.27	-43.58	-41.70
MR (best)	-133.15	-10.39	13.00	-57.52	-27.09	-25.13
MF	-55.19	-3.64	60.08	8.39	-29.99	-19.36