

ARTICLE OPEN



Predicting the inhibition efficiencies of magnesium dissolution modulators using sparse machine learning models

Elisabeth J. Schiessler¹, Tim Würger^{2,3}, Sviatlana V. Lamaka², Robert H. Meißner^{2,3}, Christian J. Cyron^{1,4}, Mikhail L. Zheludkevich^{2,5}, Christian Feiler² and Roland C. Aydin¹

The degradation behaviour of magnesium and its alloys can be tuned by small organic molecules. However, an automatic identification of effective organic additives within the vast chemical space of potential compounds needs sophisticated tools. Herein, we propose two systematic approaches of sparse feature selection for identifying molecular descriptors that are most relevant for the corrosion inhibition efficiency of chemical compounds. One is based on the classical statistical tool of analysis of variance, the other one based on random forests. We demonstrate how both can—when combined with deep neural networks—help to predict the corrosion inhibition efficiencies of chemical compounds for the magnesium alloy ZE41. In particular, we demonstrate that this framework outperforms predictions relying on a random selection of molecular descriptors. Finally, we point out how autoencoders could be used in the future to enable even more accurate automated predictions of corrosion inhibition efficiencies.

npj Computational Materials (2021)7:193; <https://doi.org/10.1038/s41524-021-00658-7>

INTRODUCTION

Magnesium (Mg) is among the most abundant elements on our planet¹ and exhibits a high potential to revolutionize light metal engineering in a large number of application fields. Key to unlocking the full potential of Mg is to control the surface reactivity characteristics of the material due to the relatively high electrochemical reactivity of Mg, where each application field imposes unique challenges. Corrosion needs to be prevented in transport applications^{2–4} (e.g., aeronautics and automotive) to ensure the integrity of the material. Medical applications (e.g., temporary, biodegradable bone implants)^{5,6} require a degradation rate tailored to a patient-specific injury to support recovery. Batteries with a Mg anode^{7,8} need a steady dissolution rate to keep the output voltage constant. Fortunately, small organic molecules exhibit great potential to control corrosion in these highly versatile application areas—due to their almost unlimited chemical space. Each service environment fundamentally changes the boundary conditions to achieve the above mentioned goals, as the small organic molecules are usually incorporated in a complex coating system for transport applications, whereas they become a solute component of the electrolyte for Mg-air batteries.

Despite impressive progress in the screening of potential additives by efficient high throughput techniques^{9–12}, experimental approaches alone cannot possibly explore more than a tiny fraction of the vast space of compounds with potentially useful properties. However, data-driven computational methods^{13–21} can explore large areas of chemical space orders of magnitude faster, and can thus be exploited to preselect promising chemicals prior to deep experimental testing. Concomitantly, computational techniques^{22–27} can be utilized to unravel the underlying chemical mechanisms of corrosion and its inhibition, which in turn provide additional input features for

predictive quantitative structure-property relationship (QSPR) models.

Naturally, data-driven methods cannot make reliable predictions for molecules outside the domain of their respective training data (e.g., for compounds that exhibit functional groups or elemental species not present in the training set). Hence, the dataset employed for training has to reflect the complexity of the relevant chemical environment, and should ideally be a large, reliable, as well as chemically diverse and balanced database to enable accurate and robust predictions for a broad range of materials. However, the versatility of the vast chemical space of interest is associated with a wide range of different functional moieties and molecular features, and renders it challenging to identify meaningful input features to develop predictive models with a wide applicability. Cheminformatics software packages like *alvaDesc*²⁸ and *RDKit*²⁹ provide a large variety of molecular descriptors ranging from structural and topological features to more complex input features like molecular signatures³⁰ and molecular fingerprints. Furthermore, advances in computing power and simulation algorithms over the last decades enabled multiscale simulations (density functional theory calculations, molecular dynamics simulations, and finite element modeling)^{25,26,31–37}, thus providing even more potentially useful molecular descriptors for the training of data-driven models^{18,38}. Additional sets of molecular descriptors might be based on properties of the used material as well as information on the service environment.

The quality of predictive models substantially depends on the selected molecular descriptors, as input features with low relevance to the target property will degrade the model. Especially the correlation (or its absence) of descriptors derived from computer simulation with the experimental performance of corrosion inhibiting agents is controversially discussed^{15,39–41}.

¹Institute of Material Systems Modeling, Helmholtz-Zentrum Hereon, Geesthacht, Germany. ²Institute of Surface Science, Helmholtz-Zentrum Hereon, Geesthacht, Germany. ³Institute of Polymers and Composites, Hamburg University of Technology, Hamburg, Germany. ⁴Institute for Continuum and Material Mechanics, Hamburg University of Technology, Hamburg, Germany. ⁵Institute for Materials Science, Faculty of Engineering, Kiel University, Kiel, Germany. ✉email: christian.feiler@hereon.de; roland.aydin@hereon.de

However, it was demonstrated that they can be highly relevant in models that combine them with input features derived from the molecular structure¹⁸. Statistical methods such as analysis of variance (ANOVA) are well established, computationally cheap tools for the identification of relevant features and parameters^{42–45}, but may struggle to capture intricate dependencies between variables. This problem can be overcome by machine learning techniques for sparse feature selection^{14,46–49}.

In this paper, we propose and compare two different sparse feature selection strategies: statistical analysis using ANOVA *f*-tests^{42–45} and recursive feature elimination based on random forests^{47,49–52}, using training data for the Mg alloy ZE41. The training data relate results of density functional theory (DFT) calculations and molecular descriptors generated by the alvaDesc cheminformatics software package to known corrosion inhibition efficiencies of chemical compounds. We demonstrate how our feature selection strategies can be combined with deep learning into a sparse, predictive QSPR (quantitative structure-property relationship) framework. Moreover, we demonstrate how in this context autoencoders^{53,54} can be used for contour maps and anomaly detection.

RESULTS AND DISCUSSION

The software package alvaDesc was utilized to generate a set of 5290 potential input features for our model. The obtained values were divided into different subcategories, ranging from counts of simple structural features of molecules to arcane descriptors derived from chemical graph theory. After removing all molecular descriptors that exhibited constant values or were essentially zero, 1254 descriptors remained and were augmented by six molecular descriptors derived from DFT calculations (*cf.* “Methods” section). In the resulting set of 1260 molecular descriptors (features), we searched for those input features with the greatest impact on the corrosion inhibition responses of 60 small organic molecules on ZE41 (target). A list of the considered molecules can be found in Supplementary Table 7, along with their SMILES strings and

experimentally determined inhibition efficiencies. We only used data of dissolution modulators from our experimental database⁵⁵ with a molecular weight of less than 250 Da that were employed at a concentration of 0.05 M.

For sparse feature selection, we applied two different approaches: The first one was based on individual feature selection via an *f*-test based analysis of variance (ANOVA) to analyse the individual importance of the different molecular descriptors. The second one was a grouped feature selection approach utilizing recursive feature elimination with random forests as the underlying regressor to analyse the importance of *n*-tuples of molecular descriptors. For a detailed description of the applied methods, *cf.* “Methods” section. We chose to look for the top 3, 5, and 63 (equivalent to 5%) most relevant features respectively, and repeated each approach multiple times to overcome any bias induced by specific random seeds. To evaluate and compare the selection methods, as well as the predictive power of the selected features, we trained several deep learning models using the identified molecular descriptors as respective sole inputs. As an additional baseline for comparison we used models trained on randomly selected features, as well as a model trained on the full dataset.

All analyses and trainings were performed with a reduced dataset, where a randomly chosen 10% of samples (*i.e.*, six samples, *cf.* Table 3) were withheld and subsequently served as a representative example of completely unknown validation data. Furthermore, a full 10-fold cross validation was performed on all deep learning models. An overview of the workflow is depicted in Fig. 1.

Individual feature selection

First, we utilized an *f*-test based ANOVA algorithm to rank each molecular descriptor according to its individual significance for predicting the inhibition efficiencies of ZE41 via its *f*-score. Features may be deemed significant if their score is $\gg 1$.

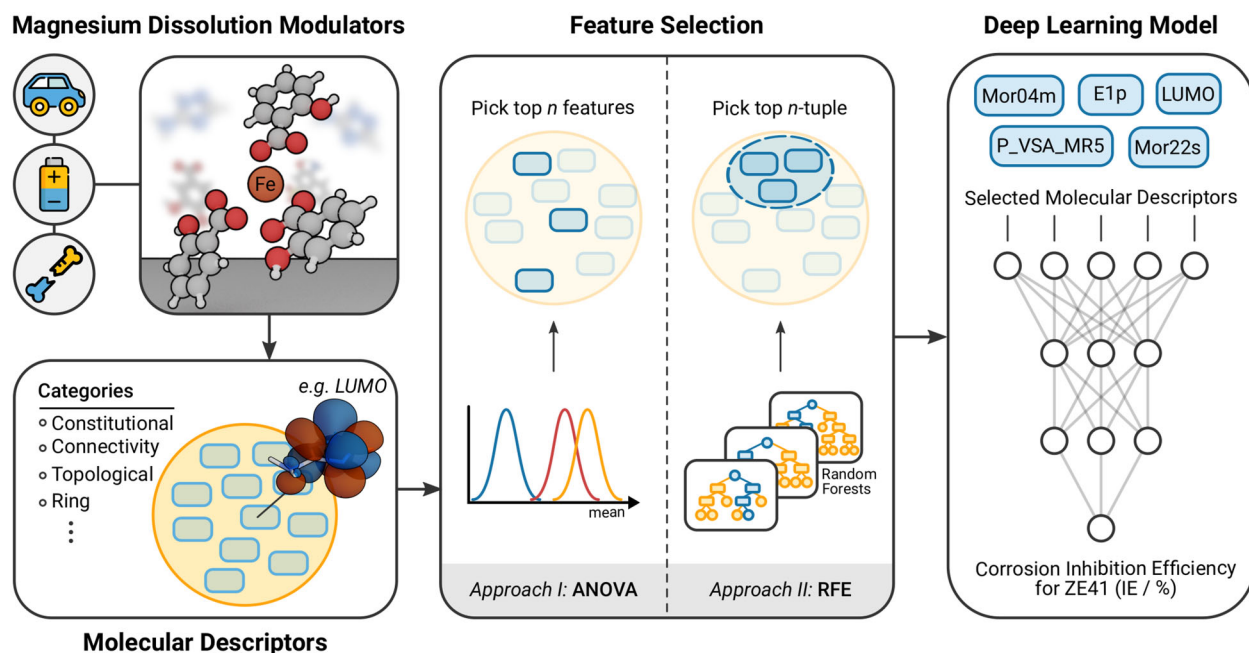


Fig. 1 Workflow overview. Our overarching objective is the prediction of the magnesium corrosion inhibition efficiency of different molecular dissolution modulators. To this end, first relevant molecular features are selected either by (approach I) analysis of variance (ANOVA) or by (approach II) recursive feature elimination based on random forests (RFE), a type of machine learning. The best-performing feature set defines the input for a deep learning model. This model allows the desired predictions of quantitative structure-property relationships (QSPR) for the efficiency of magnesium dissolution modulators, in our study specifically for the magnesium alloy ZE41.

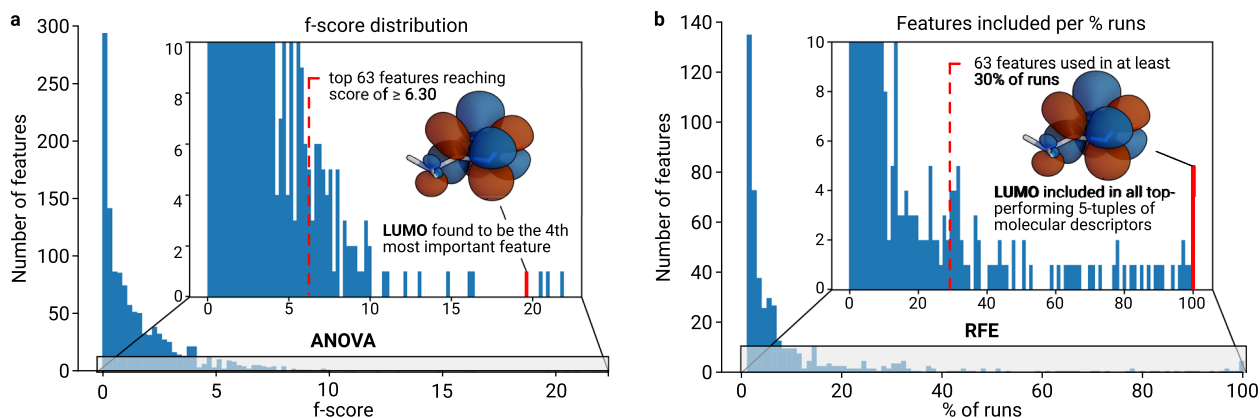


Fig. 2 Feature distributions. **a** Distribution of f -scores as calculated by ANOVA. The top 63 features reach a score of 6.3 or higher, with only 11 features scoring 10 or above. **b** The recursive feature elimination (RFE) identifies a total of 504 features over a series of 100 runs with random initialization as potential candidates for a top 63-tuple. Selecting among them those identified in at least 30% of the runs (frequency analysis) can be used to define the most relevant 63 features.

One can define the n top scoring molecular descriptors by simply ranking them via their f -score.

In our study we observed f -scores in the range from 0 to 21.92, with the vast majority of features ($\approx 92\%$) scoring below 5, *cf.* Fig. 2a. Selecting the top 3, 5, or 63 (i.e., 5%) features translates to f -score thresholds of 19.7, 16.1, and 6.3, respectively (corresponding to p -value thresholds of 0.00004, 0.00019, and 0.0155 respectively). The top five identified descriptors are CATS2D_03_AP, CATS3D_03_AP, CATS3D_02_AP, LUMO/eV, and P_VSA_MR_5 (in descending order of relevance), for the set of top 63 descriptors *cf.* Supplementary Table 2.

In particular, one of the included DFT-derived input features, the lowest unoccupied molecular orbital energy levels (LUMO), was identified as one the most relevant descriptors. Three of the five most relevant input features belong to the class of CATS descriptors^{56,57}, which are linked to properties of potential pharmacophores and are related to the discovery of novel drugs, since they indicate whether a ligand is likely to bind to a receptor site of a biological macromolecule⁵⁸. They also seem to encode structural information on functional moieties that are capable of forming coordinative bonds with ions of interest, rendering them highly relevant for the development of our model, as the inhibition efficiency of the small organic molecules is strongly dependent on their capability to form complexes with Mg^{2+} and $Fe^{2+/3+}$. The P_VSA class is comprised of 2D descriptors that reflect the sum of atomic contributions to the van-der-Waals surface area⁵⁹. The P_VSA descriptor identified by the ANOVA approach is related to the polarizability of the chemicals in our data set.

Grouped feature selection

As the interplay and correlations between parameters can have a significant impact on the quality of the prediction, it may not be sufficient to merely select the individually most predictive features and use them as the combined input for a predictive model⁴⁷. Therefore, we additionally identified the 3-tuples, 5-tuples and 63-tuples of *grouped* most relevant features via recursive feature elimination (RFE) using random forests. We performed 100 runs of RFE with varying random seeds, where a random forest consisting of 100 trees was trained in each run. Subsequently, the n -tuples that won most often were selected to be the most relevant grouped features with LUMO, P_VSA_MR_5, Mor04m (selected in 83/100 runs) for the 3-tuples and LUMO, P_VSA_MR_5, Mor04m, E1p, Mor22s (selected in 21/100 runs) for the 5-tuples. It is noteworthy, that the energy level of the lowest unoccupied molecular orbital (LUMO) of the compounds in the training set,

which was derived from DFT calculations, was again among the most relevant features, along with P_VSA_MR_5. Furthermore, different descriptors belonging to the class of 3D-MoRSE (Molecular Representation of Structures based on Electron diffraction) were selected^{60,61}. These are abbreviated as “Mor” and are a mathematical representation of XRD patterns where the obtained signals can be weighted by previously discussed schemes. E1p belongs to the class of WHIM descriptors which are 3-dimensional descriptors that collect information about size, shape, symmetry, and atom distribution of the molecule. E1p is related to the atoms distribution and density around the origin and along the first principal component axis. The index p indicates that the selected descriptor is calculated by weighting the atoms with their polarizability value.

In case of the 63-tuple, no group was found to be inherently most relevant. We therefore artificially constructed the most relevant group by a frequency analysis of all features that were included at least once in any of the RFE runs. Among these 504 features, interestingly only the ones in the top 5-tuple occurred in every single run. 135 features ($\approx 27\%$) were identified just once, and 302 ($\approx 60\%$) were found to be in at most 5 supports. The top 63 features were included in at least 30% of all runs, *cf.* Fig. 2b, for the full list *cf.* Supplementary Table 3.

This underlines that molecular descriptors derived from quantum mechanical calculations can be highly relevant input features for models that predict the corrosion inhibition efficiency of small organic molecules for Mg alloys. This is in good agreement with our findings for commercially pure Mg containing 220 ppm iron impurities (CPMg220), where the frontier orbital energy gaps exhibited moderate correlation with the corresponding inhibition efficiencies¹⁸, and could be utilized to obtain a robust predictive model in combination with structural input features. The results of others^{39–41} suggest that this type of descriptor should be taken with care because its relevance may be compromised if not combined with structural input features. Yet, as demonstrated also by our above results, if properly used, it can be a highly powerful feature for the prediction of corrosion inhibition efficiency.

To counter potential bias from the choice of the validation set when selecting the most relevant input features, we performed a 5-fold cross validation, i.e., selecting a different validation set and repeating the whole feature selection process using RFE described above independently five times. Due to computational limitations, we performed this cross-validation only on the grouped 5-tuple of features, as results obtained from subsequent deep learning models suggest an optimal trade-off between a low number of input features and a low computational cost on the one hand and

Table 1. Median statistics over the full 10-fold cross validation.

No. of features	3			5			63			1260
Model type	Tiny model			Small model			Medium model			Large model
Selection method	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	
RMSE/pp	66	56	66	57	50	66	50	50	56	63
R^2	0.40	0.51	0.30	0.68	0.66	0.41	0.60	0.62	0.50	0.35
Pearson's <i>r</i>	0.61	0.71	0.54	0.82	0.81	0.64	0.77	0.79	0.70	0.59
<i>p</i> -value	0.23	0.13	0.27	0.05	0.06	0.17	0.08	0.06	0.12	0.22

Median values of root mean squared errors (RMSE), coefficients of determination (R^2), correlation coefficients (Pearson's *r*) and *p*-values of the full 10-fold cross validation for all trained models by model type and feature selection method (a: ANOVA, b: RFE, c: random selection).

a high predictive accuracy on the other hand. The initially identified top performing 5-tuple of molecular descriptors was confirmed by this cross-validation, along with two other 5-tuples, all three of which agreeing on four out of five descriptors. The first of the three identified sets consists of LUMO, P_VSA_MR_5, Mor04m, E1p, HOMO, the second one of LUMO, P_VSA_MR_5, Mor04m, E1p, Mor22s and the third one of LUMO, P_VSA_MR_5, Mor04m, E1p, CATS3D_02_AP.

Predictive models using deep learning techniques

From the ranked list of individually most relevant features (selected by ANOVA), we used the top 3, 5, and 63 molecular descriptors to train three deep learning models, from here on called M3a (tiny model), M5a (small model), M63a (medium model). We performed a complete 10-fold cross validation, i.e., we split the dataset into ten equal parts (folds) and subsequently withheld one fold as a test set, while the rest of the data served as training set. On each fold, every model was trained 100 times with varying random seeds to obtain results largely independent on specific random initializations. Subsequently, we repeated the same procedure with the top 3, 5, and 63 most relevant molecular descriptors obtained by grouped feature selection via RFE as input for the three neural network models M3b (tiny model), M5b (small model), and M63b (medium model).

Finally, we selected 3, 5, and 63 random molecular descriptors to train three neural network models M3c (tiny model), M5c (small model), M63c (medium model) as a reference base line to assess the quality of the aforementioned models M3a, M3b, M5a, M5b, M63a, and M63b. The input features for these models were re-drawn from the set of 1260 available features in each of the 100 training runs.

As an additional baseline we trained a deep neural network M1260 (large model) which uses all available molecular descriptors as its input. This model can be considered the joint limit case of the above three feature selection methods ANOVA, RFE, and random in case that the number of selected features is increased to its maximal value of 1260.

In Table 1 we report for all the above neural network models median values (across the ten folds) of four key statistical measures of their predictive capabilities, that is, of the root mean squared error RMSE (given in percentage points), the coefficient of determination R^2 , Pearson's correlation coefficient *r*, and the *p*-value. In Table 1 we observe several consistent trends. First, all statistical measures of predictive capability noticeably improve when the number of input features is increased from 3 to 5 to 63 for all the three feature selection methods (ANOVA, RFE, or random). Second, the two sparse feature selection procedures (ANOVA and RFE) consistently outperform in all measures a simple random feature selection, which underlines their practical value. Third, the two sparse feature selection procedures (ANOVA and

RFE) exhibit a similar performance, with RFE slightly outperforming ANOVA with respect to RMSE, which can in many respects be considered the most relevant one of the four statistical measures. This underlines that grouped features selection has indeed—as one would also expect—advantages over individual features selection, though at least in the framework used herein only to a limited extent. A fourth important observation is the decline of performance when increasing the number of input features to 1260. This can be understood from the fact that such unspecific input dilutes the relevant information harbored by the input in a way that makes systematic learning of QSPR more difficult. Quite interestingly, for the two sparse features selection methods (ANOVA and RFE)—unlike for the random feature selection—the performance already stagnates when increasing the number of input features from 5 to 63, indicating that they can help to identify a very small group of features that carries nearly the whole information relevant for predictions.

It is noteworthy that even when using a sparse feature selection method, the error of the predictions based on the selected features still remains substantial. While fully overcoming this problem would go beyond the scope of this paper, we further investigated into the reasons of this problem. Analysing our data we found that the performance of predictions based on sparse feature selection is substantially adversely affected by only a few outliers. To illustrate this, we consider more closely compound no. 13, 3,5-Dinitrobenzoic acid. Unlike all the other 59 molecules in our data base, it contains an NO₂ functional group. This important chemical difference is supposedly the reason why the information carried by the other compounds cannot help a neural network to make accurate predictions also for 3,5-Dinitrobenzoic acid, which indeed results in a very large error for any of the above introduced predictive neural networks. Naturally, such a large error affects the otherwise very good performance of predictions based on sparse feature selection methods much more adversely than the generally much less accurate predictions based on randomly selected features. To demonstrate this, we show in Table 2 the results for one specific fold where we manually removed from the validation set 3,5-Dinitrobenzoic acid. Evidently, this substantially improves the predictions in particular made on the basis of grouped feature selection, while the quality of predictions based on of tiny or small sets of randomly selected features remains rather limited. Detailed information about the fold, validation and neural network predictions underlying to Table 2 are presented in Supplementary Tables 5 and 6. We performed Pearson correlation tests for all different models presented in Table 2 and observed in particular for neural networks receiving input features obtained from grouped feature selection a positive correlation coefficient of 0.97 and significant *p*-values below 0.01. Figure 3 illustrates the performance of the deep neural networks M5b and M1260 for the (reduced) validation set discussed in Table 2.

Table 2. Statistics on the representative validation set.

No. of features	3			5			63			1260
Model Type	Tiny model			Small model			Medium model			Large model
Selection method	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>c</i>	
RMSE/pp	50	24	66	51	26	60	49	23	40	38
R^2	0.56	0.94	0.53	0.54	0.94	0.95	0.58	0.94	0.94	0.95
Pearson's <i>r</i>	0.75	0.97	0.73	0.73	0.97	0.97	0.76	0.97	0.76	0.97
<i>p</i> -value	0.14	0.01	0.16	0.16	0.01	0.01	0.14	0.01	0.08	0.01

Root mean squared errors (RMSE), coefficients of determination (R^2), correlation coefficients (Pearson's *r*) and *p*-values of the representative validation set predictions for all trained models by model type and feature selection method (a: ANOVA, b: RFE, c: random selection). 3,5-Dinitrobenzoic acid (compound No. 13) was omitted for calculation of the statistical values as its molecular features are in parts substantially outside of the domain covered by the training data.

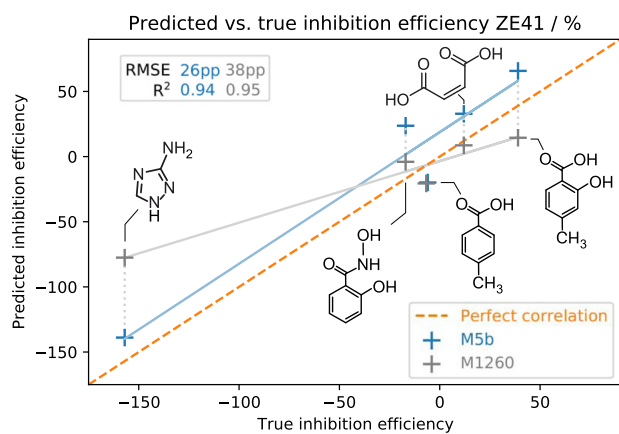


Fig. 3 Model performance. Predicted vs. true target values for the validation sets as obtained by M5b and M1260. Linear regressions are depicted as accordingly colored lines. 3,5-Dinitrobenzoic acid was excluded as it contains features that are outside of the domain of the trained model.

Comparing the median performance over all cross validation folds with that on the representative validation set showcases the potential of predictive modeling when combined with appropriate outlier detection methods. As pointed out above, a few outliers can have a drastic impact on the quality of the predictive models. In particular, one of the ten cross validation folds contains outliers that consistently yielded very poor results across all models and metrics. For this reason we elected to present median rather than mean values across all statistics, for the corresponding mean values table *cf.* Supplementary Table 4. Besides outlier detection, repeating the feature selection process for each model and each fold can also increase performance.

Autoencoders

So-called autoencoders are a type of neural network that is not used for predictions but rather to learn a lower-dimensional representation (code) of the input data, from which the original input can be reconstructed as accurately as possible (*cf.* "Methods" section). Herein we applied an autoencoder with a code of dimension 2 to the 5-tuple of features determined by grouped feature selection. The resulting two-dimensional representation of the 60 chemical compounds studied herein is plotted in Fig. 4a. Subsequently, we used the decoder part of the autoencoder to generate a contour map of predicted inhibition efficiencies across the whole two-dimensional reduced feature space, Fig. 4b to make anomalies even easier to spot with the naked eye. It is immediately noticeable as a prominent anomaly in the plot of the

reduced (two-dimensional) feature space that there are two samples with a highly negative inhibition efficiency within a cluster of samples with a (moderately) positive inhibition efficiency. The first one is 4-hydroxybenzoic acid with an inhibition efficiency of -170% whose parent system salicylic acid causes a considerably higher inhibition efficiency of 37% despite very similar molecular features. Addition of another hydroxyl group in 3,4-dihydroxybenzoic acid (the second outlier) leads to a further increase of the Mg^{2+} binding ability resulting in an inhibition efficiency of -270% . The behavior of the latter can be attributed to the significantly higher stability constant of the corresponding complex of 3,4-dihydroxybenzoic acid with Mg ($\log K(Mg^{2+}) = 9.84$) in comparison to that of salicylic acid ($\log K(Mg^{2+}) = 4.7$)^{62,63}. We assume that a similar effect is the reason for the unique behavior of 4-hydroxybenzoic acid although there is no stability constant available in the literature to support this claim. Additionally, the corresponding ligands do not only shift dissolution equilibria, but they also compete with OH^- for binding Mg^{2+} thus preventing the formation of a semi-protective $Mg(OH)_2$ layer on the substrate. Consequently, 4-hydroxybenzoic acid and 3,4-dihydroxybenzoic acid are currently investigated concerning their potential as effective additives for Mg-air battery electrolytes.

In summary, we have pointed out above how sparse feature selection methods can help to identify those molecular descriptors that carry the most valuable information for predictions of the corrosion inhibition efficiency of organic molecules on the degradation of magnesium alloys. Our results clearly demonstrate that in addition to classical structural descriptors also those directly derived from DFT calculations can be highly relevant for data-driven predictions. Interestingly, our methods of sparse feature selection reveal that the Chemically Advanced Template Search (CATS) descriptors form a particularly valuable basis for predictions. These are generally known to bear great potential for e.g., the AI-driven discovery of drugs⁶⁴. Our results suggest that the pharmacophore properties encoded therein can also help to describe the capacity of small organic molecules to form complexes with metal ions like Mg^{2+} and $Fe^{2+/3+}$. This appears natural since atoms that may act as hydrogen bond acceptors (e.g., a nitrogen atom with a lone pair) may also act as donor for the formation of a coordinative bond in another context. In some cases an intuitive understanding of the relevance of descriptors selected above may be difficult. Yet, it is striking that the DFT-derived descriptor LUMO seems to play a significant role. This claim is corroborated by the outcome of both the individual and grouped feature selection. Our above analyses were not biased in any way by any expectation of specific features becoming dominant. Yet, LUMO was selected approximately 240 times more often by our smart feature selection algorithms than expected from random probability (*cf.* Supplementary Notes) which is a strong hint at a possible causal relationship between LUMO and

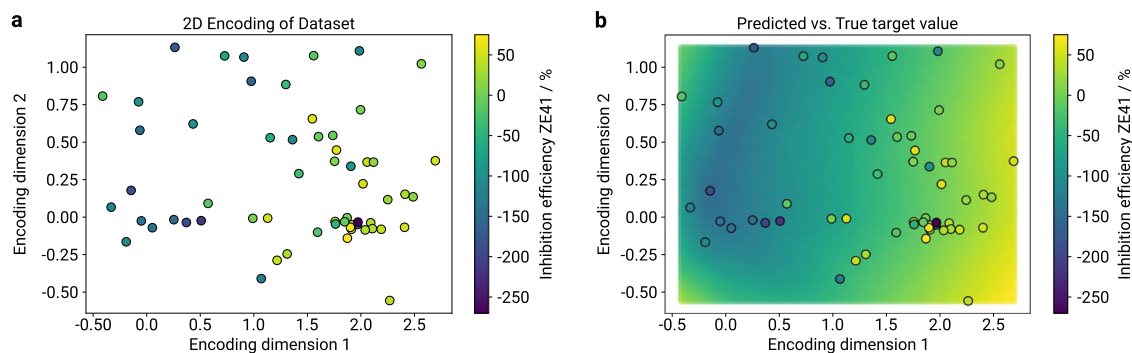


Fig. 4 Using autoencoders for outlier detection and contour maps. **a** Input features reduced to a two-dimensional code. **b** The decoder part in combination with an appropriate predictive model (such as a deep neural network) can be used to generate contour maps across the space spanned by the dimensions of the two-dimensional code.

the corrosion inhibition efficiency. Using the example of a specific fold within our 10-fold cross validation, we pointed out that the elimination or proper treatment of outliers can be expected to play a key role in further improving the accuracy of feature-based predictions of the corrosion inhibition efficiency. We showcased the ability of autoencoders to detect potential anomalies within datasets, which can be especially useful when working with small datasets. Note that the affected samples were included in all analyses and training steps as is. Yet, as apparent from the discussion above, it is very likely that the development of methods for a special treatment or at least detection of outliers could be an important step to improve data-driven predictions of corrosion inhibition efficiencies substantially, which opens up a promising avenue of future research.

METHODS

Molecular descriptor generation

To define molecular descriptors, we first determined the structures of the 60 chemical compounds of interest using the quantum chemical software package Turbomole 7.4⁶⁵ at the TPSSh/def2SVP^{66,67} level of density functional theory. Six of the molecular descriptors considered herein are directly derived from the output of the performed DFT calculations. These are the frontier orbital energies (HOMO, LUMO) as well as the frontier orbital energy gap ($\Delta E_{\text{H-L}}$), the calculated heat capacities (C_p , C_v) and the chemical potential (μ) calculated at 293 K. The thermodynamic properties were derived from the calculated vibrational frequencies using the Turbomole module *freqh* with default parameters for the calculations. The Cartesian coordinates resulting from our DFT calculations are subsequently used as input for the cheminformatics software package alvaDesc 1.0²⁸ to generate roughly 5000 molecular descriptors related to structural features. After omitting molecular descriptors with constant values and/or those that are close to zero, we used the remaining 1254 descriptors in combination with the above mentioned six DFT descriptors as input features for our sparse feature selection method.

Dataset preprocessing

We randomly selected 10% of the available data (i.e., six samples) using `scikit-learn`'s `train-test-split`⁶⁸ that are withheld from all further preprocessing, analysis, and training. These samples serve as an unknown validation set, and are used to validate the predictive abilities of the trained models. A representative validation set is illustrated in *cf.* Table 3. The index is used for numbering of the 60 chemical compounds of interest. We applied linear min-max scaling to all descriptors to map their values on the interval $[-1, 1]$. The target variable (corrosion inhibition efficiency) was mapped on the interval $[0, 1]$.

Data analysis—individual features

To identify the most relevant molecular descriptors for predicting inhibition efficiency, we considered two approaches. The first was to regard each feature individually, and determine its influence on predicting

Index	Compound description	Inhibition efficiency ZE41/%
0	3-Amino-1,2,4-triazole	−157
5	4-Methylsalicylic acid	39
13	3,5-Dinitrosalicylic acid	38
36	Maleic acid	12
45	<i>p</i> -Toluic acid	−6
54	Salicylhydrocamic acid	−17

Set aside representative validation set (randomly selected).

the target variable, i.e. to look for the individually most relevant features. We did so by means of *f*-test based analysis of variance (ANOVA)^{42–45}. An *f*-test (or *F*-test) is a test to see whether two independent, identically distributed variables X_1 and X_2 have the same variance. The *f*-score is given by $f = \sigma_{X_1}^2 / \sigma_{X_2}^2$, with $\sigma_{X_i}^2$ denoting the variance of X_i . The null hypothesis may then be rejected if *f* is either below or above a chosen threshold α .

F-test based ANOVA calculates an *f*-score for every molecular descriptor compared to the target variable (corrosion inhibition efficiency). This score provides a statistic (with an $F(1, k-2)$ -distribution, where *k* is the number of samples) for each descriptor for testing the hypothesis whether its distribution is the same one as the one of the target variable. The higher the *f*-score, the higher the presumed relevance of a descriptor. Herein, we used the top 3, 5, and 63 (i.e., 5%) descriptors as input for a subsequent deep learning framework.

Data analysis—grouped features

Those descriptors that individually hold the most amount of information need not necessarily work best together as a group when used as the input for a deep learning model. Thus we also identified *n*-tuples of features that are most relevant as a group via recursive feature elimination (RFE)⁴⁷. RFE repeatedly fits a chosen regression model, and then discards a fraction of features found to be least relevant for decision making. This process is repeated until only the desired *n* descriptors remain. As the underlying regression model, we choose random forests^{49–51}. A random forest is a so-called ensemble learning method, i.e., a collection of individual predictors, over which an average is calculated. This reduces overfitting and increases generalizability of the model, which is especially relevant when the training set is of limited size. The random forest consists of a number of decision trees, each of which only has access to a (randomly chosen) subset of all features for making the best possible prediction. The RFE algorithm is run 100 times with varying random seeds to counter statistical artifacts. Depending on the chosen *n*, one or more *n*-tuples of features may be selected by this process more often than other combinations. If this was the case, we picked the *n*-tuple selected most often. However, the larger *n* gets, the less likely this becomes. Thus, if this was not the case, we artificially composed the best *n*-tuple based on the frequency distribution of all descriptors included in any of the tuples selected at least once as most relevant *n*-tuple.

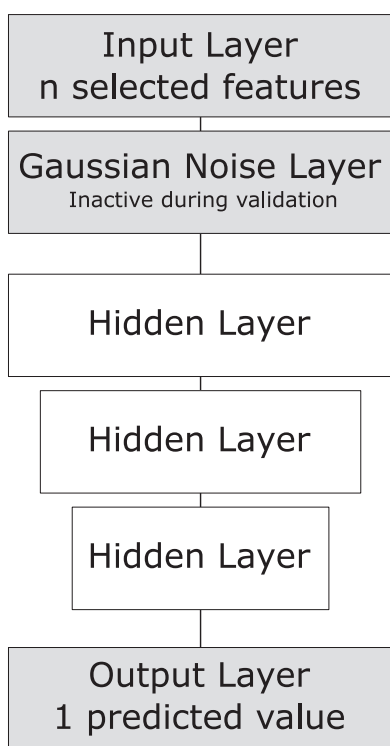


Fig. 5 Example architecture for deep learning models. General architecture of a deep learning model used for predicting corrosion inhibition efficiencies of chemical compounds (output) from molecular features (input).

Deep learning models

We evaluated the predictive value of the features identified either by *f*-test-based ANOVA or RFE by using them as input features for a deep neural network that was trained to predict the corrosion inhibition efficiency. The predictive quality of this network was then evaluated on the representative validation set withheld in the very beginning from the data (see above). Thereby we used four different types of deep neural networks: tiny models (three input features), small models (five input features), medium models (63 input features) and large models (containing the full set of 1260 available input features). Each of these models (deep neural networks) was composed by three hidden layers with a `relu` activation function (cf. Fig. 5). They were trained for 25 epochs using an Adam optimizer and the mean squared error (MSE) of the scaled target values as the loss function. Since the dataset was very small (only 54 training samples after withholding six samples for the representative validation set) the input data was first passed through a Gaussian noise layer with $\mu = 0$ and $\sigma = 0.1$ for each model. This layer added some Gaussian random noise in each epoch, which effectively served as a data augmentation technique and helped to improve generalization of the model and to reduce overfitting. The Gaussian noise layer was deactivated when predictions were made for the (previously unseen) validation data. The hyperparameters varied depending on the number of input parameters (model sizes) were the number of units in each hidden layer, as well as the learning rate for the Adam optimizer. For details the reader is referred to the supplementary material.

Autoencoders

Recently, autoencoders have attracted substantial attention in dimensionality reduction in the context of deep learning^{69–71}. Autoencoders are however not used for predictions. Rather their objective is to generate an approximation of the input data as close as possible to themselves after compressing them through a bottleneck. Autoencoders consist of three parts: an encoder that learns how to distill the most relevant information from the input; the code, i.e., the condensed information gained from the input; and lastly the decoder, which learns how to re-construct the input data as accurately as possible from the code (cf. Fig. 6).

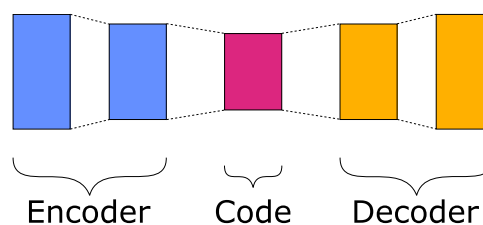


Fig. 6 Example architecture for autoencoders. Schematic illustration of an autoencoder. Each bar represents a dense layer, bars of same size indicate same number of neurons in the respective layers.

As one has substantial freedom in choosing the dimension of the code, one case use autoencoders to reduce, e.g., the 1260 input features in our problem to e.g., 2–5 key variables. Of course, the lower the dimension of the code, i.e., the greater the compression of the input data, the greater the reconstruction error typically becomes. Note that while autoencoders are quite a powerful tool for dimensionality reduction, they are not a feature selection method in the classical sense⁷¹. Rather they are similar to principal component analysis (PCA) which can also be used for dimensionality reduction. Neither the code produced by the autoencoder nor the principal components found by PCA have a direct physical correspondence to any of the input features. Instead, PCA constructs a linear projection of the input data onto a basis of the closest lower rank representation of the original data space. In general, a unique inversion of this process does not exist. Similarly, autoencoders typically learn a highly nonlinear mapping which approximates a bijection between the original and the latent data dimensions up to the reconstruction error⁷². The great advantage of autoencoders compared to PCA is that the decoder part can thus be used for predictions on generic data reconstructed from the latent space. We trained an autoencoder with a code of dimension 2, which was suitable to plot a two-dimensional representation of the chosen number of input features (for hyperparameters cf. Supplementary Table 1). Moreover, its decoder was able to map any point in this two-dimensional reduced feature space to a predicted corrosion inhibition efficiency of ZE41 (cf. Fig. 4). Note that besides providing a low-dimensional representation of the input data, autoencoders can also be used to reduce noise within a dataset, or to detect potential anomalies in the data⁷⁰.

DATA AVAILABILITY

The data used for this study is available at <https://doi.org/10.5281/zenodo.5564824>.

CODE AVAILABILITY

The code used for this study is available at <https://doi.org/10.5281/zenodo.5564824>.

Received: 1 June 2021; Accepted: 21 October 2021;

Published online: 01 December 2021

REFERENCES

- Anderson, D. L. Chemical composition of the Mantle. *J. Geophys. Res.* **88 Suppl.**, 41–52 (1983).
- Taub, A. I. & Luo, A. A. Advanced lightweight materials and manufacturing processes for automotive applications. *MRS Bull.* **40**, 1045–1053 (2015).
- Joost, W. J. & Krajewski, P. E. Towards magnesium alloys for high-volume automotive applications. *Scr. Mater.* **128**, 107–112 (2017).
- Dziubińska, A., Gontarz, A., Dziubiński, M. & Barszcz, M. The forming of magnesium alloy forgings for aircraft and automotive applications. *Adv. Sci. Tech.* **10**, 158–168 (2016).
- Luthringer, B. J. C., Feyerabend, F. & Willumeit-Römer, R. Magnesium-based implants: a mini-review. *Magnes. Res.* **27**, 142–54 (2014).
- Brar, H. S., Platt, M. O., Sarntinoranont, M., Martin, P. I. & Manuel, M. V. Magnesium as a biodegradable and bioabsorbable material for medical implants. *Jom* **61**, 31–34 (2009).
- Deng, M. et al. Ca/In micro alloying as a novel strategy to simultaneously enhance power and energy density of primary Mg-air batteries from anode aspect. *J. Power Sources* **472**, 228528 (2020).
- Zhang, T., Tao, Z. & Chen, J. Magnesium-air batteries: from principle to application. *Mater. Horiz.* **1**, 196–206 (2014).

9. Meeusen, M. et al. A complementary electrochemical approach for time-resolved evaluation of corrosion inhibitor performance. *J. Electrochem. Soc.* **166**, C3220–C3232 (2019).
10. Muster, T. H. et al. A rapid screening multi-electrode method for the evaluation of corrosion inhibitors. *Electrochim. Acta* **54**, 3402–3411 (2009).
11. White, P. A. et al. A new high-throughput method for corrosion testing. *Corros. Sci.* **58**, 327–331 (2012).
12. White, P. A. et al. Towards materials discovery: assays for screening and study of chemical interactions of novel corrosion inhibitors in solution and coatings. *N. J. Chem.* **44**, 7647–7658 (2020).
13. Chen, F. F. et al. Correlation between molecular features and electrochemical properties using an artificial neural network. *Mater. Des.* **112**, 410–418 (2016).
14. Meftahi, N. et al. Machine learning property prediction for organic photovoltaic devices. *npj Comput. Mater.* **6**, 1–8 (2020).
15. Winkler, D. A. et al. Towards chromate-free corrosion inhibitors: structure-property models for organic alternatives. *Green. Chem.* **16**, 3349–3357 (2014).
16. Le, T. C. & Winkler, D. A. Discovery and optimization of materials using evolutionary approaches. *Chem. Rev.* **116**, 6107–6132 (2016).
17. Galvão, T. L., Novell-Leruth, G., Kuznetsova, A., Tedim, J. & Gomes, J. R. Elucidating structure–property relationships in aluminum alloy corrosion inhibitors by machine learning. *J. Phys. Chem. C* **124**, 5624–5635 (2020).
18. Feiler, C. et al. In silico screening of modulators of magnesium dissolution. *Corros. Sci.* **163**, 108245 (2020).
19. Würger, T. et al. Data science based Mg corrosion engineering. *Front. Mater.* **6**, 53 (2019).
20. Würger, T. et al. Exploring structure–property relationships in magnesium dissolution modulators. *npj Mater. Degrad.* **5**, 2 (2021).
21. Zeller-Plumhoff, B. et al. Exploring key ionic interactions for magnesium degradation in simulated body fluid—a data-driven approach. *Corros. Sci.* **182**, 109272 (2021).
22. Yuwono, J. A., Taylor, C. D., Frankel, G. S., Biribilis, N. & Fajardo, S. Understanding the enhanced rates of hydrogen evolution on dissolving magnesium. *Electrochem. Commun.* **104**, 106482 (2019).
23. Milošev, I. et al. Editors' choice—The effect of anchor group and alkyl backbone chain on performance of organic compounds as corrosion inhibitors for aluminum investigated using an integrative experimental-modeling approach. *J. Electrochem. Soc.* **167**, 061509 (2020).
24. Würger, T., Feiler, C., Vonbun-Feldbauer, G. B., Zheludkevich, M. L. & Meißner, R. H. A first-principles analysis of the charge transfer in magnesium corrosion. *Sci. Rep.* **10**, 15006 (2020).
25. Feiler, C., Mei, D., Luthringer-Feyerabend, B., Lamaka, S. & Zheludkevich, M. Rational design of effective Mg degradation modulators. *Corrosion* **77**, 204–208 (2021).
26. Poberžnik, M. et al. DFT study of n-alkyl carboxylic acids on oxidized aluminum surfaces: from standalone molecules to self-assembled-monolayers. *Appl. Surf. Sci.* **525**, 146156 (2020).
27. Fockaert, L. et al. ATR-FTIR in Kretschmann configuration integrated with electrochemical cell as in situ interfacial sensitive tool to study corrosion inhibitors for magnesium substrates. *Electrochim. Acta* **345**, 136166 (2020).
28. Mauri, A. *Methods in Pharmacology and Toxicology*, 801–820 (Humana Press Inc., 2020).
29. Landrum, G. et al. Rdkit: open-source cheminformatics. <https://www.rdkit.org/> (2016).
30. Mikulskis, P., Alexander, M. R. & Winkler, D. A. Toward interpretable machine learning models for materials discovery. *Adv. Intell. Syst.* **1**, 1900045 (2019).
31. Pérez-Sánchez, G., Galvão, T. L., Tedim, J. & Gomes, J. R. A molecular dynamics framework to explore the structure and dynamics of layered double hydroxides. *Appl. Clay Sci.* **163**, 164–177 (2018).
32. Klink, S., Höche, D., La Mantia, F. & Schuhmann, W. FEM modelling of a coaxial three-electrode test cell for electrochemical impedance spectroscopy in lithium ion batteries. *J. Power Sources* **240**, 273–280 (2013).
33. Hammerich, M. et al. Heterodiazocines: synthesis and photochromic properties, trans to cis switching within the bio-optical window. *J. Am. Chem. Soc.* **138**, 13111–13114 (2016).
34. Ma, R., Huang, D., Zhang, T. & Luo, T. Determining influential descriptors for polymer chain conformation based on empirical force-fields and molecular dynamics simulations. *Chem. Phys. Lett.* **704**, 49–54 (2018).
35. Ash, J. & Fourches, D. Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *J. Chem. Inf. Model.* **57**, 1286–1299 (2017).
36. Pereira, F. & Aires-de Sousa, J. Machine learning for the prediction of molecular dipole moments obtained by density functional theory. *J. Cheminform.* **10**, 43 (2018).
37. Nørskov, J. K., Abild-Pedersen, F., Studt, F. & Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proc. Natl Acad. Sci.* **108**, 937–943 (2011).
38. Richert, C. & Huber, N. A review of experimentally informed micromechanical modeling of nanoporous metals: from structural descriptors to predictive structure–property relationships. *Materials* **13**, 3307 (2020).
39. Morales-Gil, P., Walczak, M. S., Cottis, R. A., Romero, J. M. & Lindsay, R. Corrosion inhibitor binding in an acidic medium: interaction of 2-mercaptobenzimidazole with carbon-steel in hydrochloric acid. *Corros. Sci.* **85**, 109–114 (2014).
40. Winkler, D. A. et al. Using high throughput experimental data and in silico models to discover alternatives to toxic chromate corrosion inhibitors. *Corros. Sci.* **106**, 229–235 (2016).
41. Kokalj, A. et al. Simplistic correlations between molecular electronic properties and inhibition efficiencies: Do they really exist? *Corros. Sci.* **179**, 108856 (2021).
42. Johnson, K. J. & Synovec, R. E. Pattern recognition of jet fuels: comprehensive GC × GC with anova-based feature selection and principal component analysis. *Chemom. Intell. Lab. Syst. J.* **60**, 225–237 (2002).
43. Burgard, D. R. *Chemometrics: Chemical and Sensory Data* (CRC Press, 2018).
44. Kim, T. K. Understanding one-way anova using conceptual figures. *Korean J. Anesthesiol.* **70**, 22–26 (2017).
45. Bijma, F., Jonker, M., van der Vaart, A. & Erné, R. *An Introduction to Mathematical Statistics* (Amsterdam University Press, 2017).
46. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16–28 (2014).
47. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
48. Solorio-Fernández, S., Carrasco-Ochoa, J. A. & Martínez-Trinidad, J. F. A review of unsupervised feature selection methods. *Artif. Intell. Rev.* **53**, 907–948 (2020).
49. Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, 278–282 (IEEE, 1995).
50. Genuer, R., Poggi, J.-M. & Tuleau-Malot, C. Variable selection using random forests. *Pattern Recogn. Lett.* **31**, 2225–2236 (2010).
51. Chavent, M., Genuer, R. & Saracco, J. Combining clustering of variables and feature selection using random forests. *Commun. Stat. Simul. Comput.* **50**, 426–445 (2021).
52. Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. Choosing feature selection and learning algorithms in qsar. *J. Chem. Inf. Model.* **54**, 837–843 (2014).
53. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J. & Chen, H. Application of generative autoencoder in de novo molecular design. *Mol. Inform.* **37**, 1700123 (2018).
54. Samanta, S., O'Hagan, S., Swainston, N., Roberts, T. J. & Kell, D. B. Vae-sim: a novel molecular similarity measure based on a variational autoencoder. *Molecules* **25**, 3446 (2020).
55. Lamaka, S. V. et al. Comprehensive screening of Mg corrosion inhibitors. *Corros. Sci.* **128**, 224–240 (2017).
56. Schneider, G., Neidhart, W., Giller, T. & Schmid, G. 'Scaffold-Hopping' by topological pharmacophore search: a contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896 (1999).
57. Fechner, U., Franke, L., Renner, S., Schneider, P. & Schneider, G. Comparison of correlation vector methods for ligand-based similarity searching. *J. Comput. Aid. Mol. Des.* **17**, 687–698 (2003).
58. Grisoni, F., Merk, D., Byrne, R. & Schneider, G. Scaffold-hopping from synthetic drugs by holistic molecular representation. *Sci. Rep.* **8**, 1–12 (2018).
59. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **18**, 464–477 (2000).
60. Devinyak, O., Havrylyuk, D. & Lesyk, R. 3D-MoRSE descriptors explained. *J. Mol. Graph. Model.* **54**, 194–203 (2014).
61. Schuur, J. H., Selzer, P. & Gasteiger, J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comp. Sci.* **36**, 334–344 (1996).
62. Dean, J. A. *Lange's Chemistry Handbook*. (University of Tennessee, McGrawHill, Inc, 1999).
63. Smith, R. & Martell, A. *Critical Stability Constants, Vol. 3. Other Organic Ligands*, vol. 365 (Plenum Press, 1977).
64. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **12**, 1–22 (2020).
65. TURBOMOLE. V7.4. *A Development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989–2019 since 2007* (TURBOMOLE GmbH, 2019).
66. Staroverov, V. N., Scuseria, G. E., Tao, J. & Perdew, J. P. Comparative assessment of a new nonempirical density functional: molecules and hydrogen-bonded complexes. *J. Chem. Phys.* **119**, 12129–12137 (2003).
67. Eichkorn, K., Weigend, F., Treutler, O. & Ahlrichs, R. Auxiliary basis sets for main row atoms and transition metals and their use to approximate coulomb potentials. *Theor. Chem. Acc.* **97**, 119–124 (1997).
68. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

69. Wang, Y., Yao, H. & Zhao, S. Auto-encoder based dimensionality reduction. *Neurocomputing* **184**, 232–242 (2016).
70. Sakurada, M. & Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis* 4–11 (ACM, 2014).
71. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
72. Almotiri, J., Elleithy, K. & Elleithy, A. Comparison of autoencoder and principal component analysis followed by neural network for e-learning using handwritten recognition. *2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* 1–5 (IEEE, 2017).

ACKNOWLEDGEMENTS

Funding by the Helmholtz Association is gratefully acknowledged. T.W. and C.F. gratefully acknowledge funding by the Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service) via Projektnummer 57511455. R.M. gratefully acknowledges funding by the Deutsche Forschungsgemeinschaft (D.F.G., German Research Foundation) via Projektnummer 192346071-SFB 986 and Projektnummer 390794421-GRK 2462.

AUTHOR CONTRIBUTIONS

E.J.S., T.W., S.V.L., R.H.M., C.J.C., M.L.Z., C.F., and R.C.A. contributed to the conception and design of the study. C.F. generated the molecular descriptor database. E.J.S. did the theoretical analyses and wrote the supporting code. E.J.S., T.W., R.C.A., and C.F. evaluated the quality of the presented models. E.J.S. and T.W. created the figures. E.J.S. and C.F. wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read, and approved the submitted version.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00658-7>.

Correspondence and requests for materials should be addressed to Christian Feiler or Roland C. Aydin.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021