Predicting the cytotoxicity of MXenes - layered materials - using machine learning methods.

Maciej Marchwiany
Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
Magdalena Birowska (Magdalena.Birowska@fuw.edu.pl)
University of Warsaw https://orcid.org/0000-0001-6357-7913
Mariusz Popielski
Uniwersytet Warszawski Wydzial Fizyki
Jacek A. Majewski
Uniwersytet Warszawski Wydzial Fizyki
Agnieszka M. Jastrzębska
Politechnika Warszawska Wydzial Inzynierii Materialowej

Research

Keywords: MXene, machine learning, layered materials, cytotoxicity

Posted Date: February 14th, 2020

DOI: https://doi.org/10.21203/rs.2.23632/v1

Predicting the cytotoxicity of MXenes - layered materials - using machine learning methods

Maciej Marchwiany,*[a] Magdalena Birowska,[b], Mariusz Popielski,[b] Jacek A. Majewski [b] and Agnieszka M. Jastrzębska[c]

Background : Prediction of the compound cytotoxicity is a crucial issue in the development of new drugs and potential biomedical applications. Experimental studies are time-consuming and expensive. Machine learning models can quickly predict the cytotoxicity of compounds, by extracting new insights from large materials and biological data sets, and provide further guidance for experimental studies.

Results : Here, we identify the most relevant features that are responsible for the cytotoxic behavior of layered MXenes materials. The most important result of our work is the identification of 2D MXenes specific surface parameters as responsible for the potential cytotoxicity of these materials, in particular, the presence of transition metal oxides and Lithium atoms on the surface. After successful verification of the correct predictions of our model, we have also succeeded in predicting toxicity for 2D MXenes not tested *in vitro*. Hence, we have been able to complement the existing knowledge coming from *in vitro* studies.

Conclusions : Our results allow for the future selection of synthesis methods preventing surface oxidation, which should allow production of non-toxic 2D MXenes. Such materials might find application in many fields of science and technology, especially in biotechnology and nanomedicine.

Keywords: MXene, machine learning, layered materials, cytotoxicity

Background

MXenes are defined as early transition metal carbides, nitrides, and carbonitrides. They have received much attention as their unique 2D crystal structure can be easily tuned to produce dramatic improvement in material properties[1]. Therefore, it is not surprising that they have been successfully applied among most important fields of materials science and technology[2]. The term 'MXene' reflects the unique two-dimensional (2D) structure of the material in which the formula $M_{n+1}X_nT_z$ perfectly matches the arrangement of its layered features. The atomically thin layers are attached to each other via a weak so called van der Waals forces. In this formula M is early transition metal, X is carbon and/or nitrogen, n=1, 2, or 3, whereas T_z corresponds to the surface terminal functional groups (e.g., -OH, =O, -F)[3]. The family of MXenes has expanded rapidly since the discovery of their first representative – the $Ti_3C_2T_z$ phase in 2011 by Naguib *et al.*[4].

It should be noted, that the first package of MXenes included only several phases with 19+ successfully synthesized in subsequent years[5]. Since that time, only a few years have passed and now researchers are able to predict new MXene phases theoretically[6], and new phases have been successfully synthesized (see the review paper[7]). As can be seen, many more of them are yet to be obtained for further solution processing and potentially valuable applications in industry. Hence, their development has risen exponentially. On the other hand, the experimental pursuit is far away from the theoretical peloton. Such booming development typically forces such large pursuit in research, that safety verification by systematic and in-depth studies pose a great challenge. This aspect especially affects the quality of biological studies in which toxicological studies can be misleading and difficult to verify, in view of incomplete material characterization. The best examples of such findings are the antibacterial properties that were claimed for the 2D MXenes. Initial results showed biocidal properties of $Ti_3C_2T_z$[8] together with the assumed nano-blade mechanism of action[9], and a lack of this effect for Ti_2CT_z[10]. But in-depth studies in relation to material properties and structure finally showed a lack of antibacterial properties of the pristine $Ti_3C_2T_z$[11]. In the case of MXenes cytotoxicity, the first studies concerned *in vitro* testing of multilayered $Ti_3C_2T_z$ MXene and showed a potential threat related to the generation of reactive oxygen species (ROS)[12]. Further studies showed differences in toxicological effects in view of MXenes stoichiometry (i.e., $Ti_3C_2T_z$ or Ti_2CT_z)[13]. Also, the importance of flake thickness was highlighted by us not only in the case of material stability but also potential toxicity[14]. We could even say that right now we are at the moment, where some new solutions are needed to extract the most promising representatives of MXenes with the highest potential for application and the lowest cytotoxicological threats. It becomes obvious that it is impossible to carry out screening investigations for all MXenes phases in reasonable time, and low costs. The most time- and cost-consuming analyses are undoubtedly the biological studies, which are also inevitable to push through MXenes applications in industry. What is more, many certification procedures involve verification of the safety of market products containing the claimed nanomaterials. Accordingly, there is a strong demand for theoretical solutions that could overcome the problem of so many complicated analyses.

One such solution is the machine learning (ML). Machine learning has so far proved its applicability for cytotoxicity studies of

[a] *Address, Interdisciplinary Centre for Mathematical and Computational Modelling (ICM), University of Warsaw, Pawińskiego 5a, 02-106 Warsaw, Poland; E-mail: M.Marchwiany@icm.edu.pl*

[b] *Address, University of Warsaw, Faculty of Physics, 00-092 Warsaw, Pasteura 5, Poland; E-mail: Magdalena.Birowska@fuw.edu.pl, Mariusz.Popielski@fuw.edu.pl, jacek.majewski@fuw.edu.pl*

[c] *Address, Warsaw University of Technology, Faculty of Materials Science and Engineering, 02-507 Warsaw, Wołoska 141, Poland; E-mail: agnieszka.jastrzebska@pw.edu.pl*

large number of various chemicals (see review paper [15]), and also recent predictions of synthesis of various MXenes compounds [16]. Thus, we claim that it can be one of best solutions for reducing the number of toxicological studies needed, and allows for minimizing failures in future biological applications. Machine learning studies concerning toxicity of drug, molecules have been carried out previously, by using deep learning and XGBoost [17] approaches, and using Atomic Fingerprints [18–23]. However, to the best of our knowledge, there is a lack of toxicological research based on ML methods concerning layered materials, in particular, 2D structures. Thus, we undertake the challenge of predicting the cytotoxicity of experimentally synthesized MXene compounds.

Our aim is to build a model based on machine learning methods, in order to predict the cytotoxicity of MXenes materials with some elemental information provided from experiments. We consider up to 71 records in a dataset, including our own experimental results as well as data collected from recently published experimental studies. The elemental information about the materials such as surface characteristics, morphology, and structure has been taken into account. Next, the ML approach has been applied in the form of Random Forest [24], which enables us to identify the most important key features, that have a decisive impact on MXenes cytotoxicity. Then, we apply Principle Component Analysis (PCA) [25] as feature engineering to improve our model. We used the key features to train machine learning models. The models were checked by a 10-fold cross-validation scheme. We use this model to predict the cytotoxicity of 19 experimentally examined MXene compounds.

Results and discussion

The toxicological *in vitro* data for 2D MXenes have been taken from recently published high-throughput screening experiments, therefore, they are reliable and convenient for comparison. In order to determine a good quality model, it is crucial to identify a ML algorithm suited for given a dataset. We have decided to test three datatsets (two distinct ones and one which consists of these two) and various algorithms, in order to determine an appropriate model with high accuracy of prediction. The datasets used in the present study are listed below:

- Dataset I (experimental set) - Experimental data on the toxicity of 2D MXenes compounds (see Table 4). An adequate package of 2D MXene structures was selected for ML analysis based on two key criteria. The first criterion was the availability of an experimental 2D structure and a detailed description of the experiment and possible surface modification contained in the given work. The second criterion was the exact and in-depth characterization of morphology, chemical composition, and structure of the resulting 2D MXenes as well as their effect on cells *in vitro*. On this basis, it was possible to obtain comprehensive experimental data necessary to perform ML analysis. It consists of 71 records and elemental features listed in Table 5 in Appendix A.

- Dataset II (theoretical set) - data taken from the two-dimensional database [26] concerning the geometry information about the known 2D MXenes compounds (61 records). The elemental features are collected in Table 6.

- Dataset III (combined set) - dataset consists of both Dataset I and Dataset II. The elemental features are collected in Tables 5 and 6.

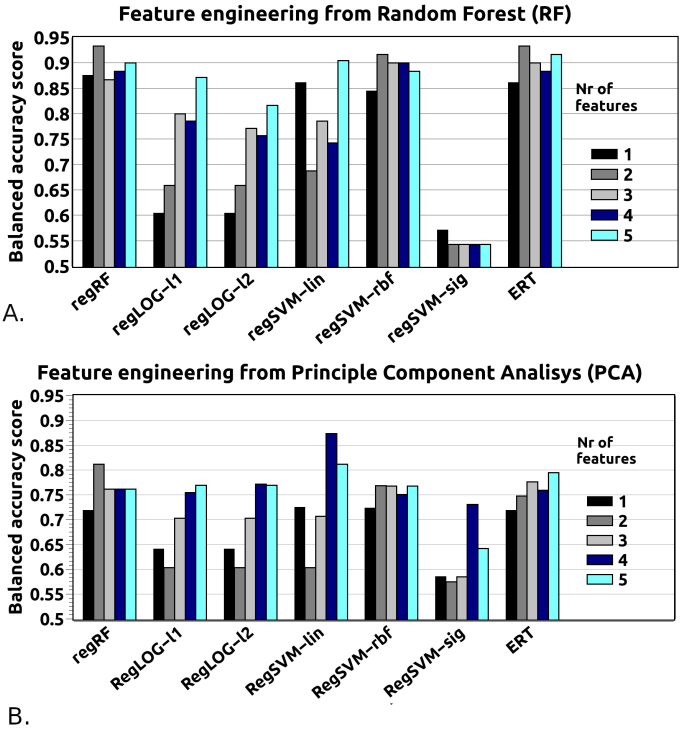

Fig. 1 Feature engineering for dataset I obtained for two methods: (A) Random Forest (RF) and (B) Principle Component Analysis (PCA).

Data analysis has been carried out for each of the datasets separately. It is worth mentioning, that the first two datasets overlapped partially, namely, the geometry of each of the compounds in the first set is known, but only 8 compounds have specific toxicity in the second set. The size and the type of the variables in the dataset determine the appropriate ML algorithm. We have used a variety of different methods well suited for the minimal size of the dataset and briefly described in **Materials and Methods** section. Detailed analysis of the applicability of machine learning algorithms can be found elsewhere [27]. Then, the models are simplified, by selecting the most important features based on Random Forest algorithm or by the construction of new features from the given ones by the use of the Principle Component Analysis (PCA). The models are tested by 10-fold cross-validation, with the performance measured by class balanced accuracy score of correct predictions [28]. The accuracy score metric is defined in the range of [0,1].

Dataset I

Our theoretical investigations (see Table 1) reveal that the accuracy score for balanced data shows a good level of precision, greater than 0.72 (except for a Support Vector Machine (SVM) with sigmoid kernel regSVM-sig) for all of the algorithms employed in this paper. Moreover, note that data balancing tech-

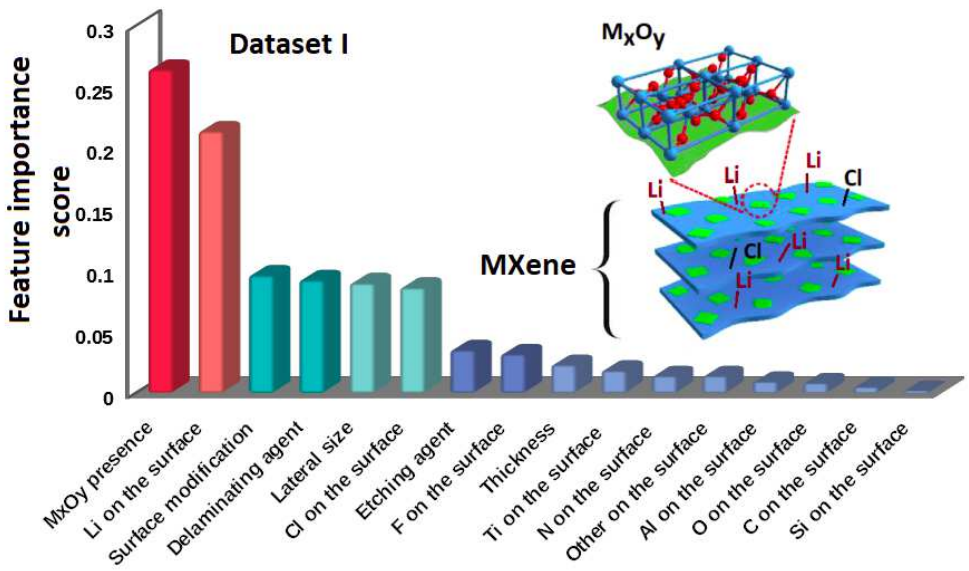

Fig. 2 Ranking of feature importance obtained from RF. The most important are two descriptors: the presence of M_xO_y and Li on the surface, whereas the next four: surface modification, delaminating agent, lateral size, and Cl on the surface are equally important with smaller weight than the previous two. All feature labels are described in Table 5.

niques improve the results approximately by a few percent (see the results for balanced versus unbalanced data collected in Table 1). The largest values are obtained in case of the SVM with rbf kernel (regSVN-rbf) for Weight and SMOTE techniques, and are equal to 0.92 and 0.93 respectively. The high accuracy obtained in case of regRF and regSVM-rbf manifest the non-linearity of the studied problem and the need for using non-parametric models. Unfortunately, this has a negative effect on understanding of the results and the underlying phenomenon. It is worth noting, that most of the variables used in this study are categorical variables described by one-hot-encoded or labeling methods. By use of the SMOTE data balancing algorithm, the results do not include pure features but contain mixed ones, which results in losing the physical interpretation of the outcomes. Thus, we have decided to use the Weight data balancing algorithms further in this study.

Table 1 The accuracy score of correct predictions are obtained for unbalanced and balanced data, for the various algorithms employed in this paper. The most accurate results are obtained for balanced data by use of the Support Vector Machine with rbf kernel (regSVM-rbf).

		balanced data:	
ML algorithms	unbalanced data	Weight	SMOTE
regRF	0.747	0.826	0.833
regLOG-l1	0.700	0.776	0.783
regLOG-l2	0.720	0.798	0.783
regSVM-lin	0.867	0.725	0.808
regSVM-rbf	0.662	0.917	0.933
regSVM-sigmoid	0.555	0.543	0.4042
ERT	0.722	0.776	0.750

In order to understand the cytotoxicity issue and its dependency on selected descriptors, feature selection and feature engineering techniques are applied further in this study. The crucial premise of feature selection is that the data contains some variables that are either redundant or irrelevant, and can thus be removed without much loss of information. Feature selection and feature engineering techniques allow simplification of the model, and hence, they can facilitate the interpretation of the studied phenomenon. In order to select the most important features, the feature importance score is obtained from a random forest analysis. Feature importance shows how much weight the model assigns to the given descriptor during predictions, and thus, gives insight into which variables are crucial for predicting the cytotoxicity of MXenes materials. For comparison purposes, we also use feature engineering from PCA.

The results reveal that three features are already sufficient for a good level of prediction accuracy, for all employed algorithms except regSVM-sig (see Figure 1). Moreover, all the models based on PCA show lower accuracy than RF, which means that there is a low correlation between the features. The PCA approach is based on correlation between the features, while feature importance from RF selects the most important, original variables. The difference in these results denotes that the correlation between features is irrelevant, and the underlying physics is written in the original variables. In addition, the feature importance score shows that there are two crucial features, four equally important features, and the rest seem to be unimportant from the RF analysis (see Figure 2). The most important features are the presence of the M_xO_y and the Li atoms on the MXenes surfaces.

Dataset II

Here we have tested the dataset with structural information of the compounds included, namely position and type of atoms. We want to note here that only part of the dataset is labelled as toxic or non-toxic, which does not allow for the use of classification methods. In order to effectively elucidate the information contained in this dataset, clustering technique has been applied. There are many methods available such as Atom-Centered Sym-

metry Functions (ACSF)[29], Coulomb Matrix[30], or Ewald Sum Matrix[31], which convert the atomic positions into variables that can be used in machine learning. We have used the Weighted Atom-Centered Symmetry Functions (wACSF)[32] as descriptors, in order to substantially decrease the number of variables. The parameters of this model have been adopted from Ref.[29].

Subsequently, the most popular clustering algorithms have been used. We have also examined the internal structure of the dataset, and then, grouped the compounds as toxic or non-toxic. All tested methods predict all compounds as non-toxic. The results reveal that the clustering technique cannot be viewed as a mechanism for toxicity prediction of MXenes. In another approach, we used abnormality detection algorithms. One-class SVM[33], Isolation forest[34] and Robust covariance[35] methods were used.

All the results presented in this subsection reveal that taking into account purely theoretical information about MXenes materials is not sufficient to build a good quality ML model with high accuracy for cytotoxicity prediction. However, this theoretical data set can be used as a component of an enlarged experimental data set I. Such combined dataset is a subject of study in the next subsection.

Dataset III

Here we present the results for the extended dataset I with variables related to the geometry of the studied structures (dataset II), to determine whether such a combined database improves model predictions.

ML algorithms	Model selection: Weight
regRF	0.845
regLOG-l1	0.845
regLOG-l2	0.727
regSVM-lin	0.781
regSVM-rbf	0.876
regSVM-sigmoid	0.545
ERT	0.793

Table 2 The metric of accuracy score of correct predictions is obtained, for various of algorithms employed in this paper.

Our study reveals that including the information related to the geometry of the compounds (see Table 2) does not improve the results, and gives a similar level of accuracy of prediction as obtained for dataset I (see Table 1), for all of the methods employed here.

Then, we have built models that include from one to five features by use of the Random Forest algorithm and PCA method, similarly to the approach for dataset I. The analysis demonstrates, that two features are sufficient to describe the toxicity of MXenes compounds with high accuracy of predictions (see Figure 3). From the feature importance ranking (see Figure 4) we find out that the topmost descriptors are the presence of M_xO_y, Li on the surface of MXenes, and surface modification with external compounds. Note, that the order of the six top important features is the same as in the case for dataset I. Aforementioned results indicate that experimental data are sufficient to build a good model, and the inclusion of theoretical information about MXenes does not qualitatively change the ML results.

Model predictions

We have searched the available literature covering the MXenes compounds for which all the elemental features listed in Table 5 (obviously without toxic feature information) have been provided, but with no *in vitro* studies. Despite the fact that there are around a hundred phases synthesized so far, we have only found 19 MXenes compounds, for which comprehensive data on the material are available (see Table 3). Therefore, data are taken from recently published high-throughput experiments.

The ML models predict two of 2D MXenes can exhibit cytotoxic properties with a high probability of prediction equal to 0.9, while the rest of them are predicted to be non-toxic (see Table 3). It is worth mentioning that for the non-toxic ones, no presence of M_xO_y on the surface has been reported. The presence of M_xO_y is the key toxicity-generating feature obtained from our studies. However, the prediction has to be viewed with caution, knowing that traditional k-fold cross-validation is highly optimistic when evaluating machine learning models, due to the fact that materials datasets are rarely uniformly distributed.

Our results indicate that knowledge about the surface and its modification are crucial issues concerning the toxicity of these layered 2D materials, whereas geometrical descriptors have little impact on the outcomes. It should be stressed that this conclusion is much more definitive than we expected at the beginning of our studies. The reason is that the chemical diversity and inhomogeneity of MXenes are already widely known and pose a major challenge in such complex analysis. The second corresponding

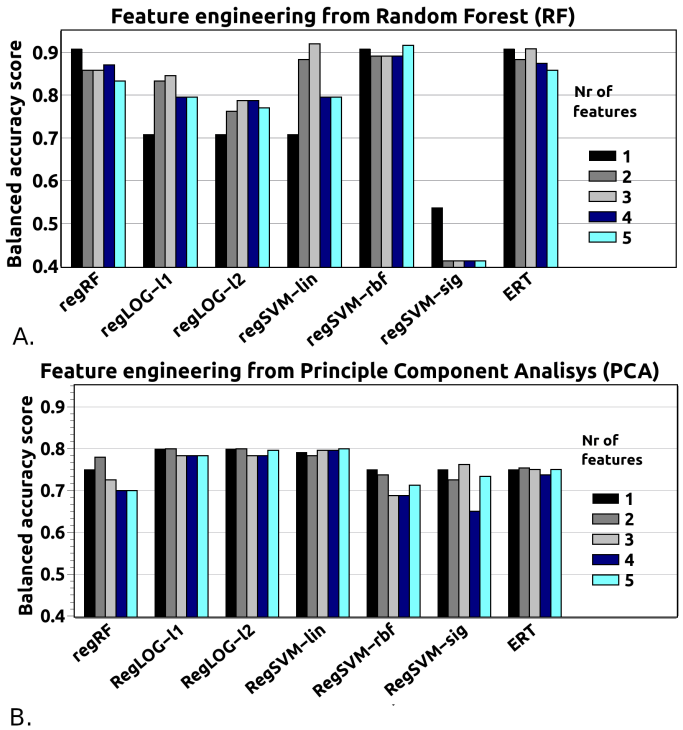

Fig. 3 Feature engineering for dataset III obtained for two methods: (A) Random Forest (RF) and (B) Principle Component Analysis (PCA).

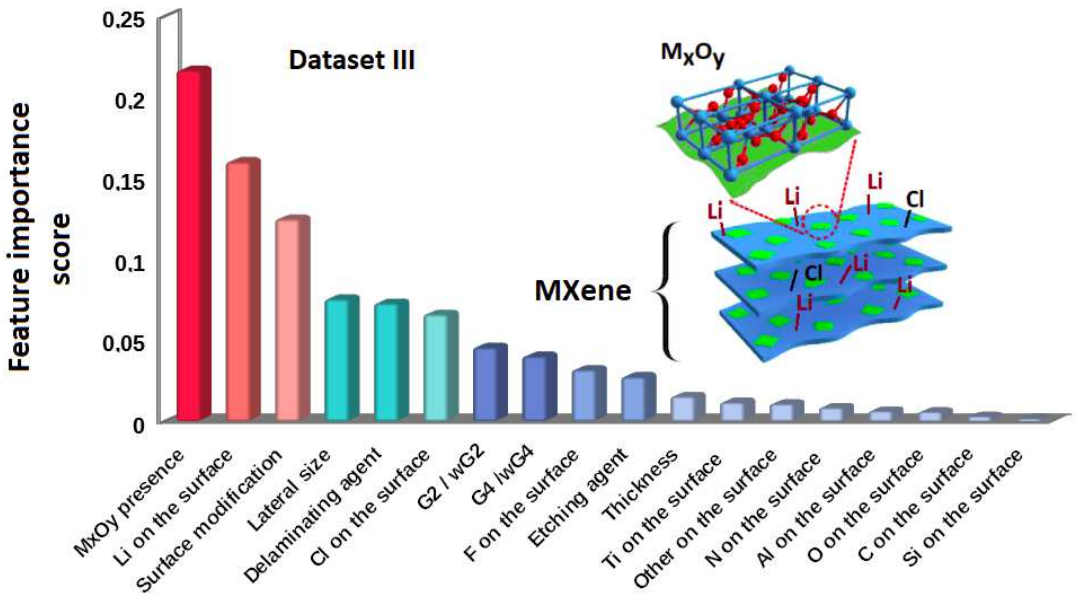

Fig. 4 Ranking of feature importance for dataset III. The most important are the first three descriptors, namely M_xO_y, Li on the surface and surface modifications, respectively. All feature labels are described in Tables 5 and 6.

Table 3 Predicted probability of cytotoxicity score of various MXenes compounds. For each of the compounds different chemical treatment, as well as chemical composition on the surface have been reported. Thus, different probability scores have been obtained. All data taken from high-throughput screening experiments (see last column). The presence of M_xO_y and Li atoms are listed in third and fourth columns, respectively. The abbreviation Us. denotes Ultrasounds, whereas the other labels are explained in table 5.

MXene [Ref.]:	Probability	M_xO_y	Li	synthesis procedure	Surface modifications
Ti_3C_2 [36]	0.18	No	Yes	LiF/HCl; Us.	No
Ti_3C_2 [37]	0.05	No	No	HF	Au
Ti_3C_2 [38]	0.18	No	Yes	LiF/HCl; Us.	No
Ti_3C_2 [39]	0.10	No	No	HF; Us.	No
Ti_3C_2 [40]	0.10	No	No	HF; Us.	No
Ti_3C_2 [41]	0.06	No	Yes	LiF/HCl	APTES+CEA
Ti_3C_2 [42]	0.07	No	Yes	LiF/HCl; Us.	DNA, Pt, Pd
Ti_3C_2 [43]	0.05	No	No	HF; Us.	Ag
Ti_3C_2 [44]	0.07	No	Yes	LiF/HCl	L-ACC
Ti_3C_2 [45]	0.07	No	Yes	LiF/HCl; Us.	PAS
Ti_3C_2 [46]	0.87	Yes	No	LiF/HCl	No
Ti_3C_2 [47]	0.88	Yes	No	LiF/HCl	No
V_2C [45]	0.05	No	Yes	LiF/HCl; Us.	PAS
V_2C [48]	0.05	No	No	NaF/HCl	No
Nb_2C [49]	0.04	No	No	NaF/HCl	No
Nb_2C [50]	0.04	No	No	HF; 60^o	No
Ti_2N [14]	0.05	No	No	KF/HCl; Us.	No
$Mo_{1.33}C$ [51]	0.04	No	No	HF; TBAOH	No
Ti_4N_3 [52]	0.03	No	Yes	KF/LiF/NaF, TBAOH	No

aspect is the divided surface characteristics. The primary strictly depends on the starting materials (MAX phases). The latter is undoubtedly far more problematic if it comes into interactions with highly sensitive systems such as biological ones. Basically, the chemical composition of the surface of MXenes almost certainly is closely related to the type of 'M' element and a resulting chemical composition of the M_xO_y passivation layer that occurs as a result of M reaction with oxygen and/or water[53]. In fact, every surface of the MXene exposed to the air can naturally react with oxygen because the freshly exposed metallic surface is energetically unsaturated and possesses high reactivity. This can happen immediately after the delamination process (but certainly must also depend on the MXene stability). What's more, the freshly exposed surface of the MXene also acquires bonding with products of chemical reactions that occur during acidic etching of the 'A' element from the MAX phase. As can be seen, the aforementioned surface-related features at first influence the material itself, but finally, may result in different biological effects such as the appearance or lack of cytotoxicity.

CONCLUSIONS

Here we present the first theoretical study concerning the toxicological aspects of 2D MXene materials. We have employed various machine learning models to study the problem of toxicity of

MXenes materials, in order to provide information that can accelerate time consuming and expensive cytotoxical experimental studies, by reducing the large number of compounds, and hence, speed up potential future applications.

Our work has demonstrated that the most important features responsible for the toxicological properties are related to the presence of transition metal oxides M_xO_y and Lithium atoms on the surface, as well as surface modification with external compounds. The presence of these features provides the guidance for the need for further expensive *in vitro* and then *in vivo* experimental studies. In other words, our detailed analysis reveals that the crucial issue is what happens on the surface, while the structural information of the systems have minimal impact on cytotoxicological aspects of MXenes materials. These results are in the line with recent experimental findings concerning the presence of Ti_xO_y on the surface [12,54,55], and the biological knowledge of cytotoxicity mechanisms [56], as well as physical and chemical intuition.

One of the goals of these studies was also to complement existing experimental studies, for which no cytotoxicological measurements have been carried out. Thus, we have predicted the cytotoxicity of 19 MXenes compounds reported widely in literature. Our studies reveal that two of them are predicted to be cytotoxic with 0.9 probability, for which the presence of the M_xO_y have been reported. The rest of the compounds are predicted to be non-toxic. Non-toxic compounds can be applied in many technological areas. Several review papers summarized their high application potential as transparent conductive films, electromagnetic interference absorption and shielding, energy storage [2], non-lithium-ion batteries, supercapacitors [57]. In addition, the potential applications in detection and sensing [58] as well as biomedical applications [59] have been widely reported.

To sum up, our methods represent a breakthrough in prediction of MXenes potential toxicity and pave the way for future applications of these materials in industry. In addition, the theoretical research methodology based on ML models presented here can be applied to other types of 2D materials exhibiting complex structure and diverse surface characteristics, such as e.g. novel 2D transition metal borides, so called MBenes [60].

METHODS

Here, we present briefly the ML algorithms used in the present study:

- Logistic regression [61] with regularization L_1 and L_2 (regLOG-L_1, regLOG-L_2). This approach allows avoiding over-learning a model even for a large number of variables. The algorithm removes unimportant features for the model.

- Random Forest (RF) [24] is commonly used for a small dataset, and must be used with care regarding over-learning. It allows for selecting the most important features.

- Support Vector Machine (SVM) [62] uses only part of the dataset, thus, it can be easily applied to a small size of dataset. The key point of prediction in the SVM algorithm is the choice of kernel. In this study, we have tested the commonly used kernels such as: linear, rbf, and sigmoid, denoted by us regSVM-lin, regSVM-rbf, regSVN-sig, respectively, throughout this paper.

- Extreme Random Tree (ERT) [63] is an extension of a Random Forest algorithm, and is known to be computationally faster than RF. Both ERT and RF are known to work well for any dataset.

Parametric models such as linear regression are used to help us understand a phenomenon by determining the functional dependences. In the case of non-parametric models such as Random Forest, the crucial issue is to identify the importance of features, and thus, it allows us to understand the studied phenomenon. Note, that other commonly used ML methods such as Kernel Ridge Regression (KKR) [64] or Neural Networks (NN) [65], are well suited for large datasets, thus, are not applicable in our case.

Table 4 Detailed information about the types of delaminated 2D MXenes compounds used in this study.

MXenes compound	
Ti_3C_2	Refs. [12,36–47,66–75]
Ta_4C_3	Refs. [76–78]
Nb_2C	Refs. [49,50,79,80]
Ti_2C	Ref. [59]
Mo_2C	Refs. [81]
$Mo_{1.33}C$	Ref. [51]
Nb_4C_3	Refs. under publications
V_2C	Refs. [48,82]
Ti_2N	Ref. [14]
Ti_4N_3	Ref. [52]

In addition, our datasets face a commonly known issue, namely, the class imbalance problem. Significantly, this problem is widely reported for the toxicity of many other materials, where the size of the positive data (toxic samples) is considerably smaller than the negative data (non-toxic samples) (see the review [83]). To solve this problem, we made use of various data-balancing algorithms such as: Weight classifier (Weight), and generating synthetic samples (SMOTE). The other commonly used algorithms such as *oversample minority class* or *undersample majority class* are not applicable in the case of MXenes materials, due to the small number of toxic records for which proper statistics cannot be built.

We have used the Python programming language with the scikit-learn [84] and XRT [85] libraries for data analysis and machine learning. The Pandas [86] library was adopted in order to read and process the data, whereas the NumPy package [87] was used to construct the features.

Experimental data and elemental features

It is well known, that ML methodology needs a significant amount of data to be reliable, and successful in toxicity predictions. On the other hand, the knowledge regarding biological properties, and especially, toxicity of the MXenes [59] has now become rich enough to obtain accurate and balanced results of predictions. Here we present detailed information about the high-throughput experimental data listed in Table 4, and the elemental features taken into account in the machine learning predictions, collected in Tables 5 and 6, respectively.

Table 5 Detailed description of the elemental features used in ML scheme and applied for the dataset I and dataset III.

Elemental feature	Description
Surface modification with external compounds	PVP, SP,MnO_x+SP, HA, Fe_xO_y+SP, PEI, PEG, DNA+Pt+Pd, Ag, L-ACC, PAS CTAC+PEG+SiO_2+c(RGDyC)+SiO_2, Au+Fe_3O_4, Au+PEG, PLL, APTES+CEA, PVA, Au
Lateral size	from few to hundredths of *nm*, from few to hundredths of μm
Thickness	from few to tens of *nm*
Etching agent	HF, LiF+HCl, LiF+HCl+$AlCl_3$, NaF+HCl, KF+HCl, KF+LiF+NaF,
Delaminating agent	Ultrasounds, DMF+high pressure+high Temp., high pressure+high Temp., TBAOH, TBAOH+ultrasounds, TPAOH, TMAOH, DMSO+ultrasounds, no additional treatment
Carbon (C) on a surface	1-Yes, 0-No
Oxygen (O) on a surface	1-Yes, 0-No
Fluor (F) on a surface	1-Yes, 0-No
Aluminium (Al) on a surface	1-Yes, 0-No
Titanium (Ti) on a surface	1-Yes, 0-No
Nitrogen (N) on a surface	1-Yes, 0-No
Chloride (Cl) on a surface	1-Yes, 0-No
Silicon (Si) on a surface	1-Yes, 0-No
Lithium (Li) on a surface	1-Yes, 0-No
Other on a surface	1-Yes, 0-No
presence of M_xO_y	1-Yes, 0-No
Toxic	1-Yes, 0-No

The abbreviations used in the table indicate: SP - Soybean phospholipid; PVP – polyvinylpyrrolidone; HA - hyaluronic acid; PEI – polyethylene imine; PEG – polyethylene glycol; PVA – polyvinyl alcohol; NMP - N-methyl-2-pyrrolidone; TMAOH - tetramethylammonium hydroxide; TBAOH - tetrabuthylammonium hydroxide; TPAOH - tetrapropylammonium hydroxide; DMF - dimethylformamide; DMSO – dimethyl sulfoxide; CTAC - cetanecyltrimethylammonium chloride; c(RGDyC) - cyclic arginine-glycine-aspartic pentapeptide; PLL – poly-L-lysine; APTES - (3-aminopropyl)triethoxysilane; CEA - carcinoembryonic antigen; L-ACC – L-ascorbic acid; PAS – polyanionic salts.

Table 6 Detailed description of the elemental features used in ML scheme for the dataset II and dataset III.

Elemental feature	Description
G2\wG2	see Ref.[29]; R_C cutoff =6, number of the functions N=10,
G4\wG4	see Ref.[29]; R_C cutoff =6, number of the functions N=10, $\lambda=\{-1,1\}$, $\zeta=\{1,2,4,16\}$

Declaration

Acknowledgements

Access to computing facilities of PL-Grid Polish Infrastructure for Supporting Computational Science in the European Research Space and of the Interdisciplinary Center of Modeling (ICM), University of Warsaw is gratefully acknowledged. We made use of computing facilities of TU Dresden ZIH within the project "TransPheMat".

Funding

The study was accomplished thanks to the funds allotted by the National Science Centre on the basis of decision no. DEC-2017/26/E/ST8/01073, within the framework of the research project 'SONATA BIS 7' no. UMO-2017/26/E/ST8/01073. M.B. is funded by the National Science Centre, Poland grant no. UMO-2016/23/D/ST3/03446. J.A.M acknowledges the support of NCN through the grant OPUS-16 (UMO-2018/31/B/ST3/03758).

Authors' contributions

MM developed the software based on ML algorithms, build the theoretical models and carried out the theoretical predictions; MB coordinated and supervised the theoretical results, prepared tables and figures, and wrote manuscript; MP collected the theoretical data, analyzed and visualized the results; JAM corrected the manuscript; AMJ designed the study and supervised the whole research as a project leader, collected the experimental data as well as coordinated the preparation of the manuscript. All authors read and approved the final manuscript.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Ethics approval and consent to participate

Authors approved ethics in publishing and consent to participate thereof.

Consent for publication

Authors declare consent for publication.

Competing interests

The authors declare that they have no competing interests.

Notes and references

1 Z. Jian-Feng, C. Hui-Yang and W. Hong-Bing, *Journal of Inorganic Materials*, 2017, **32**, 561.

2 V. M. Hong Ng, H. Huang, K. Zhou, P. S. Lee, W. Que, J. Z. Xu and L. B. Kong, *J. Mater. Chem. A*, 2017, **5**, 3039–3068.

3 L. Verger, C. Xu, V. Natu, H.-M. Cheng, W. Ren and M. W. Barsoum, *Curr. Opin. Solid State Mater. Sci.*, 2019, **23**, 149–163.

4 M. Naguib, O. Mashtalir, J. Carle, V. Presser, J. Lu, L. Hultman, Y. Gogotsi and M. W. Barsoum, *ACS Nano*, 2012, **6**, 1322–1331.

5 M. Naguib, J. Halim, J. Lu, K. M. Cook, L. Hultman, Y. Gogotsi and M. W. Barsoum, *Journal of the American Chemical Society*, 2013, **135**, 15966–15969.

6 M. Khazaei, M. Arai, T. Sasaki, C.-Y. Chung, N. S. Venkataramanan, M. Estili, Y. Sakka and Y. Kawazoe, *Advanced Functional Materials*, 2013, **23**, 2185–2192.

7 B. Anasori, M. R. Lukatskaya and Y. Gogotsi, *Nature Reviews Materials*, 2017, **2**, year.

8 K. Rasool, M. Helal, A. Ali, C. E. Ren, Y. Gogotsi and K. A. Mahmoud, *ACS Nano*, 2016, **10**, 3674–3684.

9 A. Arabi Shamsabadi, M. Sharifian Gh., B. Anasori and M. Soroush, *ACS Sustainable Chemistry & Engineering*, 2018, **6**, 16586–16596.

10 A. A. Jastrzębska, E. Karwowska, D. Basiak, A. Zawada, W. Ziemkowska, T. Wojciechowski, D. Jakubowska and A. Olszyna, *Int. J. Electrochem. Sci.*, 2017, **12**, 2159–2172.

11 A. Rozmysłowska-Wojciechowska, E. Karwowska, P. .S, T. Wojciechowski, L. Chlubny, A. Olszyna, W. Ziemkowska and A. Jastrzębska, *RSC Adv.*, 2019, **9**, 4092–4105.

12 A. Jastrzębska, A. Szuplewska, T. Wojciechowski, M. Chudy, W. Ziemkowska, L. Chlubny, A. Rozmysłowska and A. Olszyna, *Journal of Hazardous Materials*, 2017, **339**, 1 – 8.

13 A. M. Jastrzębska, E. Karwowska, T. Wojciechowski, W. Ziemkowska, A. Rozmysłowska, L. Chlubny and A. R. Olszyna, *Journal of Materials Engineering and Performance*, 2018, **28**, 1272–1277.

14 B. Soundiraraju and B. K. George, *ACS Nano*, 2017, **11**, 8892–8900.

15 Z. Yin, H. Ai, L. Zhang, G. Ren, Y. Wang, Q. Zhao and H. Liu, *Journal of Applied Toxicology*, 2019, **39**, 1366–1377.

16 N. C. Frey, J. Wang, G. I. Vega Bellido, B. Anasori, Y. Gogotsi and V. B. Shenoy, *ACS Nano*, 2019, **13**, 3031–3041.

17 L. Breiman, *Technical Report 486, University of California, Berkeley*", 1997.

18 Y. Wu and G. Wang, *International Journal of Molecular Sciences*, 2018, **19**, 2358.

19 I. Grenet, Y. Yin, J.-P. Comet and E. Gelenbe, author, 2018, pp. 335–345.

20 A. Mayr, G. Klambauer, T. Unterthiner and S. Hochreiter, *Frontiers in Environmental Science*, 2016, **3**, 80.

21 T. Lei, H. Sun, Y. Kang, F. Zhu, H. Liu, W. Zhou, Z. Wang, D. Li, Y. Li and T. Hou, *Molecular Pharmaceutics*, 2017, **14**, 3935–3953.

22 I. Grenet, Y. Yin and J.-P. Comet, *Sensors*, 2018, **18**, year.

23 I. Grenet, K. Merlo, J.-P. Comet, R. Tertiaux, D. Rouquié and F. Dayan, *Journal of Chemical Information and Modeling*, 2019, **59**, 1486–1496.

24 T. K. Ho, *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, 14-16 August 1995, **Montreal**, 278.

25 I. Jolliffe, *Principal Component Analysis, Second Edition*, Springer, 1986.

26 K. Persson, *The Materials Project*, `https://materialsproject.org/`.

27 G. Forman and I. Cohen, Knowledge Discovery in Databases: PKDD 2004, Berlin, Heidelberg, 2004, pp. 161–172.

28 L. Mosley, *QA balanced approach to the multi-class imbalance problem*, Graduate Theses and Dissertations, Iowa State University, 2013.

29 J. Behler, *The Journal of Chemical Physics*, 2011, **134**, 074106.

30 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

31 F. Faber, A. Lindmaa, O. A. von Lilienfeld and R. Armiento, *International Journal of Quantum Chemistry*, 2015, **115**, 1094–1101.

32 M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi and P. Marquetand, *The Journal of Chemical Physics*, 2018, **148**, 241709.

33 B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola and R. C. Williamson, *Neural Comput.*, 2001, **13**, 1443–1471.

34 F. T. Liu, K. M. Ting and Z. hua Zhou, In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. IEEE Computer Society, pp. 413–422.

35 W. J. D. Haan and A. T. Levin, *A Practitioner's Guide to Robust Covariance Matrix Estimation*, National Bureau of Economic Research Working Paper 197, 1996.

36 C. E. Ren, K. B. Hatzell, M. Alhabeb, Z. Ling, K. A. Mahmoud and Y. Gogotsi, *The Journal of Physical Chemistry Letters*, 2015, **6**, 4026–4031.

37 R. B. Rakhi, P. Nayak, C. Xia and H. N. Alshareef, *Scientific Reports*, 2016, **6**, 36422.

38 A. Shahzad, K. Rasool, W. Miran, M. Nawaz, J. Jang, K. A. Mahmoud and D. S. Lee, *ACS Sustainable Chemistry & Engineering*, 2017, **5**, 11481–11488.

39 L. Wu, X. Lu, Dhanjai, Z.-S. Wu, Y. Dong, X. Wang, S. Zheng and J. Chen, *Biosensors and Bioelectronics*, 2018, **107**, 69 – 75.

40 A. Arabi Shamsabadi, M. Sharifian Gh., B. Anasori and M. Soroush, *ACS Sustainable Chemistry & Engineering*, 2018, **6**, 16586–16596.

41 S. Kumar, Y. Lei, N. H. Alshareef, M. Quevedo-Lopez and K. N. Salama, *Biosensors and Bioelectronics*, 2018, **121**, 243 – 249.

42 J. Zheng, B. Wang, A. Ding, B. Weng and J. Chen, *Journal of Electroanalytical Chemistry*, 2018, **816**, 189 – 194.

43 R. P. Pandey, K. Rasool, V. E. Madhavan, B. Aïssa, Y. Gogotsi and K. A. Mahmoud, *J. Mater. Chem. A*, 2018, **6**, 3522–3533.

44 X. Zhao, A. Vashisth, E. Prehn, W. Sun, S. A. Shah, T. Habib, Y. C. Z. Tan, J. L. Lutkenhaus, M. M. Radovic and M. J. Green, *Matter*, 2019, **1**, 513.

45 V. Natu, J. L. Hart, M. Sokol, H. Chiang, M. L. Taheri and M. W. Barsoum, *Angewandte Chemie International Edition*, 2019, **58**, 12655–12660.

46 J. Zhu, Y. Tang, C. Yang, F. Wang and M. Cao, *Journal of The Electrochemical Society*, 2016, **163**, A785–A791.

47 C. J. Zhang, S. Pinilla, N. McEvoy, C. P. Cullen, B. Anasori, E. Long, S.-H. Park, A. Seral-Ascaso, A. Shmeliov, D. Krishnan, C. Morant, X. Liu, G. S. Duesberg, Y. Gogotsi and V. Nicolosi, *Chemistry of Materials*, 2017, **29**, 4848–4856.

48 F.Liu, J. Zhou, S. Wang, B. Wang, C. Shen, L. Wang, Q. Hu, Q. Huang and A. Zhou, *J. Electrochem. Soc.*, 2017, **164**, 709–A713.

49 O. Mashtalir, M. R. Lukatskaya, M.-Q. Zhao, M. W. Barsoum and Y. Gogotsi, *Advanced Materials*, 2015, **27**, 3501–3506.

50 C. Peng, P. Wei, X. Chen, Y. Zhang, F. Zhu, Y. Cao, H. Wang, H. Yu and F. Peng, *Ceramics International*, 2018, **44**, 18886 – 18893.

51 Q. Tao, M. Dahlqvist, J. Lu, S. Kota, R. Meshkian, J. Halim, J. Palisaitis, L. Hultman, M. W. Barsoum, P. O. Persson and J. Rosen, *Nature Communications*, 2017, **8**, year.

52 P. Urbankowski, B. Anasori, T. Makaryan, D. Er, S. Kota, P. L. Walsh, M. Zhao, V. B. Shenoy, M. W. Barsoum and Y. Gogotsi, *Nanoscale*, 2016, **8**, 11385–11391.

53 T. Habib, X. Zhao, S. A. Shah, Y. Chen, W. Sun, H. An, J. L. Lutkenhaus, M. Radovic and M. J. Green, *npj 2D Materials and Applications*, 2019, **3**, year.

54 A. Szuplewska, A. Rozmysłowska-Wojciechowska, S. Poźniak, T. Wojciechowski, M. Birowska, M. Popielski, M. Chudy, W. Ziemkowska, L. Chlubny, A. Moszczyńska, D. Olszyna, J. A. Majewski and A. M. Jastrzębska, *Journal of Nanobiotechnology*, 2019, **17**, 1–14.

55 A. M. Jastrzębska, A. Szuplewska, A. Rozmysłowska-Wojciechowska, M. Chudy, A. Olszyna, M. Birowska, M. Popielski, J. A. Majewski, B. Scheibe, V. Natu and M. Barsoum, *2D Materials*, 2020.

56 L. Franqui, L. de Luna, T. Loret, D. Martinez and C. Bussy, *Assessing the Adverse Effects of Two-Dimensional Materials Using Cell Culture-Based Models.*, Springer, 2019.

57 X. Li, C. Wang, Y. Cao and G. Wang, *Chemistry - An Asian Journal*, 2018, **13**,.

58 J. Zhu, E. Ha, G. Zhao, Y. Zhou, D. Huang, G. Yue, L. Hu, N. Sun, Y. Wang, L. Y. S. Lee, C. Xu, K.-Y. Wong, D. Astruc and P. Zhao, *Coordination Chemistry Reviews*, 2017, **352**, 306–327.

59 A. Szuplewska, D. Kulpińska, A. Dybko, A. M. Jastrzębska, T. Wojciechowski, A. Rozmysłowska, M. Chudy, I. Grabowska-Jadach, W. Ziemkowska, Z. Brzózka and A. Olszyna, *Materials Science and Engineering: C*, 2019, **98**, 874 – 886.

60 M. Khazaei, J. Wang, M. Estili, A. Ranjbar, S. Suehara, M. Arai, K. Esfarjani and S. Yunoki, *Nanoscale*, 2019, **11**, 11305–11314.

61 C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

62 B. E. Boser, I. M. Guyon and V. N. Vapnik, Proceedings of the Fifth Annual Workshop on Computational Learning Theory, New York, NY, USA, 1992, pp. 144–152.

63 P. Geurts, D. Ernst and L. Wehenkel, *Machine Learning*, 2006, **63**, 3–42.

64 Y. Zhang, J. Duchi and M. Wainwright, Proceedings of the 26th Annual Conference on Learning Theory, Princeton, NJ, USA, 2013, pp. 592–617.

65 W. S. McCulloch and W. Pitts, *The bulletin of mathematical biophysics*, 1943, **5**, 115–133.

66 Q. Xue, H. Zhang, M. Zhu, Z. Pei, H. Li, Z. Wang, Y. Huang, Y. Huang, Q. Deng, J. Zhou, S. Du, Q. Huang and C. Zhi, *Advanced Materials*, 2017, **29**, 1604847.

67 X. Yu, X. Cai, H. Cui, S.-W. Lee, X.-F. Yu and B. Liu, *Nanoscale*, 2017, **9**, 17859–17864.

68 L. Zhou, F. Wu, J. Yu, Q. Deng, F. Zhang and G. Wang, *Carbon*, 2017, **118**, 50 – 57.

69 H. Lin, X. Wang, L. Yu, Y. Chen and J. Shi, *Nano Letters*, 2017, **17**, 384–391.

70 C. Dai, H. Lin, G. Xu, Z. Liu, R. Wu and Y. Chen, *Chemistry of Materials*, 2017, **29**, 8637–8652.

71 G. Liu, J. Zou, Q. Tang, X. Yang, Y. Zhang, Q. Zhang, W. Huang, P. Chen, J. Shao and X. Dong, *ACS Applied Materials & Interfaces*, 2017, **9**, 40077–40086.

72 X. Chen, X. Sun, W. Xu, G. Pan, D. Zhou, J. Zhu, H. Wang, X. Bai, B. Dong and H. Song, *Nanoscale*, 2018, **10**, 1111–1118.

73 X. Han, J. Huang, H. Lin, Z. Wang, P. Li and Y. Chen, *Advanced Healthcare Materials*, 2018, **7**, 1701394.

74 E. A. Hussein, M. M. Zagho, B. R. Rizeq, N. N. Younes, G. Pintus, K. A. Mahmoud, G. K. Nasrallah and A. A. Elzatahry, *International Journal of Nanomedicine*, 2019, **14**, 4529—4539.

75 W. Tang, Z. Dong, R. Zhang, X. Yi, K. Yang, M. Jin, C. Yuan, Z. Xiao, Z. Liu and L. Cheng, *ACS Nano*, 2019, **13**, 284–294.

76 C. Dai, Y. Chen, X. Jing, L. Xiang, D. Yang, H. Lin, Z. Liu, X. Han and R. Wu, *ACS Nano*, 2017, **11**, 12696–12712.

77 H. Lin, Y. Wang, S. Gao, Y. Chen and J. Shi, *Advanced Materials*, 2018, **30**, 1703284.

78 Z. Liu, H. Lin, M. Zhao, C. Dai, S. Zhang, W. Peng and Y. Chen, *Theranostics*, 2018, **8**, 1648–1664.

79 H. Lin, S. Gao, C. Dai, Y. Chen and J. Shi, *Journal of the American Chemical Society*, 2017, **139**, 16235–16247.

80 X. Han, X. Jing, D. Yang, H. Lin, Z. Wang, H. Ran, P. Li and Y. Chen, *Theranostics*, 2018, **8**, 4491–4508.

81 W. Feng, R. Wang, Y. Zhou, L. Ding, X. Gao, B. Zhou, P. Hu and Y. Chen, *Advanced Functional Materials*, 2019, **29**, 1901942.

82 V. Natu, J. L. Hart, M. Sokol, H. Chiang, M. L. Taheri and M. W. Barsoum, *Angewandte Chemie International Edition*, 2019, **58**, 12655–12660.

83 Y. Sun, A. K. C. Wong and M. S. Kamel, *International Journal of Pattern Recognition and Artificial Intelligence*, 2009, **23**, 687.

84 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

85 K. Klementiev and R. Chernikov, Advances in Computational Methods for X-Ray Optics III, 2014, pp. 60 – 75.

86 https://pandas.pydata.org/.

87 *https://numpy.org/*.

Figures

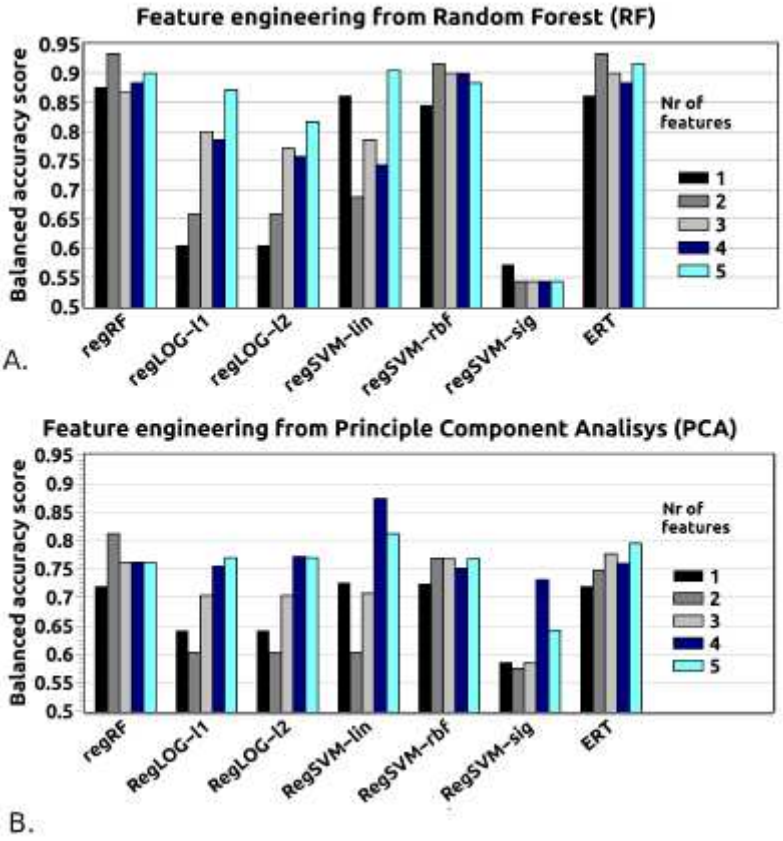

Figure 1

Feature engineering for dataset I obtained for two methods: (A) Random Forest (RF) and (B) Principle Component Analysis (PCA).

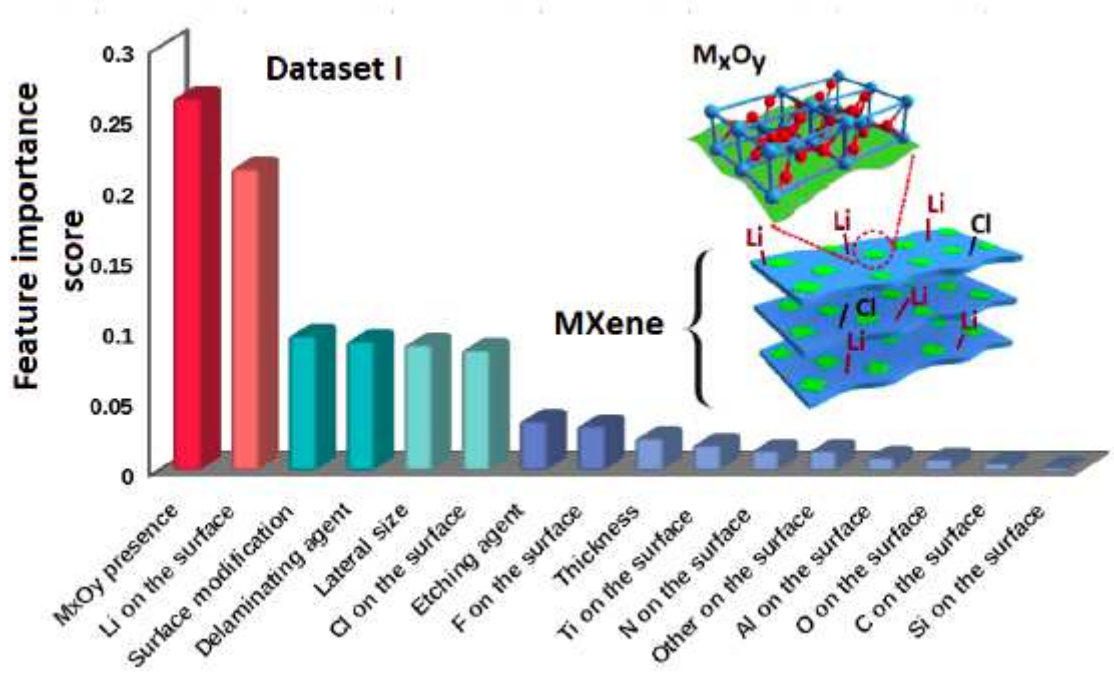

Figure 2

Ranking of feature importance obtained from RF. The most important are two descriptors: the presence of MxOy and Li on the surface, whereas the next four: surface modification, delaminating agent, lateral size, and Cl on the surface are equally important with smaller weight than the previous two. All feature labels are described in Table 5.

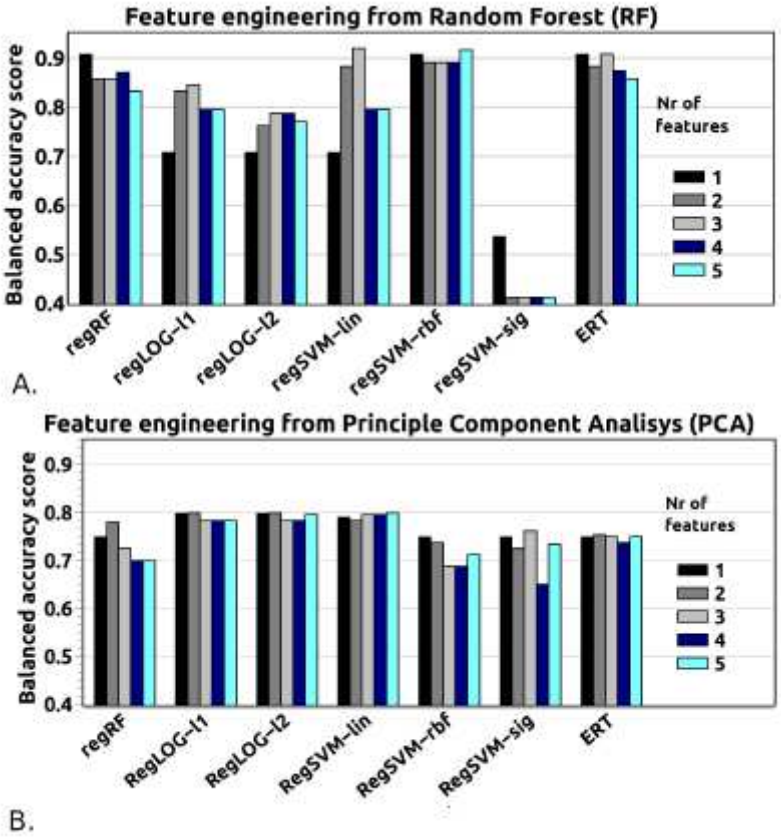

Figure 3

Feature engineering for dataset III obtained for two methods: (A) Random Forest (RF) and (B) Principle Component Analysis (PCA).

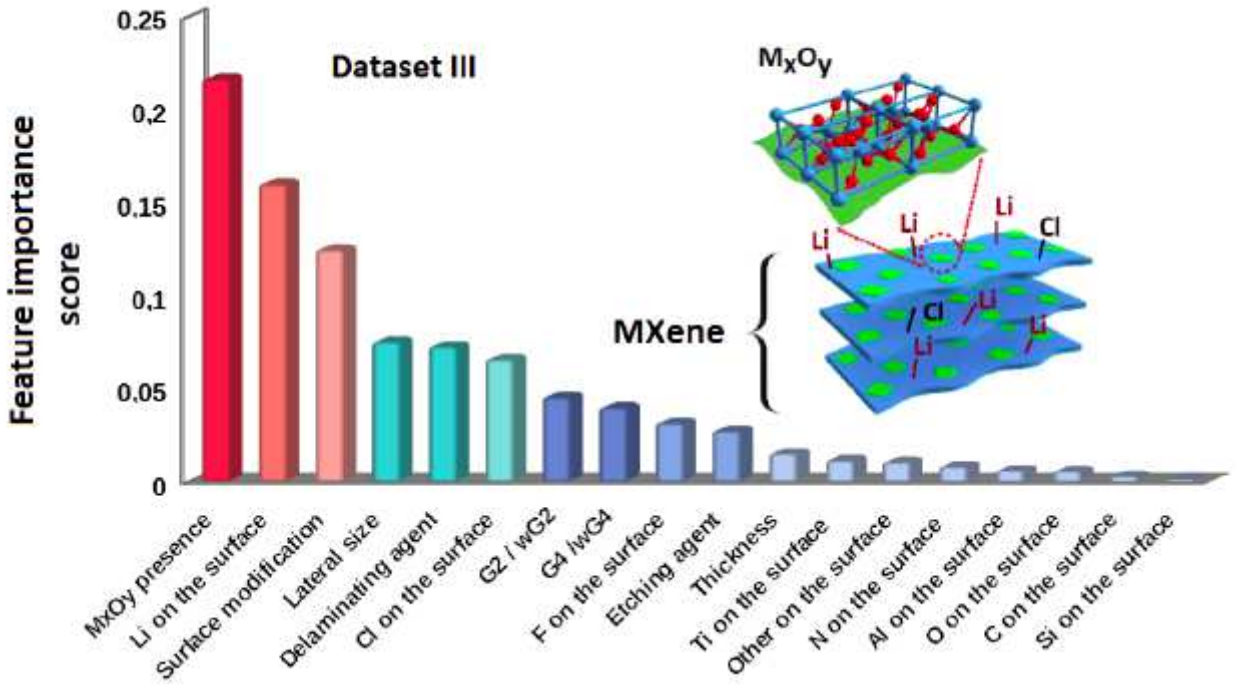

Figure 4

Ranking of feature importance for dataset III. The most important are the first three descriptors, namely MxOy, Li on the surface and surface modifications, respectively. All feature labels are described in Tables 5 and 6.