Physica C: Superconductivity and its applications 605 (2023) 1354209

Contents lists available at ScienceDirect

Physica C: Superconductivity and its applications

journal homepage: www.elsevier.com/locate/physc

Predicting the superconducting critical temperature in transition metal carbides and nitrides using machine learning

Houssam Metni [a,*], Ichiro Takeuchi [b], Valentin Stanev [b]

[a] ECPM, Université de Strasbourg, 25 Rue Becquerel, Strasbourg, 67087, France
[b] Department of Materials Science and Engineering, University of Maryland, College Park, 20742, MD, USA

ARTICLE INFO

Dataset link: https://github.com/hmetni/Superconducting_TM_C_N

Keywords:
Machine learning
Superconductivity
Transition metal carbides
Transition metal nitrides

ABSTRACT

Transition metal carbides and nitrides have unique mechanical and chemical characteristics. At low temperatures many of them also exhibit superconductivity, which can be controlled by substitutions into both the transition metal and carbon/nitrogen sites. To investigate the factors governing the superconducting state, we apply machine learning methods. We collected a dataset containing 147 materials, which was used to create a pipeline for predicting their superconducting critical temperature. When this pipeline is applied to a randomly selected test set, it shows a good performance, with R^2 of 0.82 and RMSE of 1.9 K. To explore the limits of the machine learning approach, we also use it to predict entire substitution series within the dataset. This represents a realistic test for the predictive models, which can be extremely useful when applied to new substitutions in materials systems. The performance of the pipeline in this case is much more uneven, with good predictions for some series, while for others the model shows minimal predictive power. We discuss possible reasons for these results, as well as methods to estimate the performance of machine learning on new substitution series.

1. Introduction

Transition metal (TM) carbides and nitrides have long attracted the attention of physicists and materials scientists due to their unique physical and chemical characteristics [1]. These materials possess unparalleled mechanical properties, combining very high melting points, refractory character, and chemical stability [2,3]; they are also highly conductive (of heat and electricity), and possess catalytic activities [4,5]. These materials have many traditional applications (such as protective coating for machining equipment) [6], and the combination of high conductivity with mechanical and thermal stability also makes them attractive for new applications such as energy storage and conversion electronics, photonics, and plasmonics [7–9].

Some TM carbides and nitrides are also superconducting at low temperatures, and the properties of their superconducting states have been actively investigated for decades (see, for example, Refs. [10–17]). Despite this, the exact mechanism of superconductivity in these materials is yet to conclusively established, although it likely originates from electron–phonon interactions [18–22]. This situation is far from unique to TM carbides and nitrides. While multiple classes of superconducting materials are known, it is often hard to generalize the physical and chemical features that underpin their superconducting properties. In spite of the intense research efforts (spanning in some cases decades of focused studies), the mechanisms of superconductivity of many materials families are not fully understood. These notably include both classes of ambient pressure high-temperature superconductors (cuprates [23,24] and iron-based materials [25]). Recently, researchers have turned to data-driven methods to help them address this gap and find meaningful patterns in the information collected by traditional experimental and computational methods (for a brief review of this field, see, for example, Ref. [26]). The appearance of large publicly available databases of various calculated and measured materials properties are facilitating the proliferation of studies utilizing such methods [27–32].

The data-driven approaches relying on methods such as Machine Learning (ML) allow researchers to efficiently explore a large space of existing and potentially stable materials by fully utilizing already collected information. These methods have shown great promise, and in some cases are able not only to model the properties of known superconductors, but also to suggest likely new superconducting materials for experimental exploration [27,30,31]. These successes notwithstanding, many of the proposed machine learning models fail to generalize across different types of superconductors [33]. Models trained on a single superconducting family typically have no predictive power on other groups of superconductors (see Fig. 4 in Ref. [27]). This is not

* Corresponding author.
E-mail addresses: houssam.metni@etu.unistra.fr (H. Metni), takeuchi@umd.edu (I. Takeuchi), vstanev@umd.edu (V. Stanev).

https://doi.org/10.1016/j.physc.2023.1354209
Received 13 October 2022; Received in revised form 6 January 2023; Accepted 12 January 2023
Available online 20 January 2023

a surprise — the importance of a given physical factor depends on the class of materials or the elements present. For example, there is no reason to expect similar factors to be important in determining the properties of BCS and unconventional superconductors. To address this issue, a careful examination of the ability of ML models to predict properties of groups of materials not well represented in their training set is needed. This, in turn, requires testing strategies going beyond the commonly used random train-test split of existing datasets.

In this paper, we use ML to examine the factors controlling superconductivity in TM carbides, nitrides, and carbonitrides. It contains three significant contributions to the field of applying data-driven methods for predicting superconducting properties of materials. First, we compiled a unique dataset of TM carbides and nitrides by extracting from several sources 147 reports of these materials which include their superconducting properties. Since all these materials share the same cubic NaCl structure, a single lattice parameter can describe all the unit cells in the dataset. To complement this structural descriptor, we generated a large set of elemental features. The combined dataset, consisting of 147 entries, each described by 133 features, was used to create an ML pipeline for predicting the critical temperature of the superconducting state. The pipeline shows good performance on a random subset of the data, with an R^2 higher than 0.8. This, however, represents only a relatively light test for this approach. To evaluate it further we constructed test sets representing entire substitution series. This novel testing approach, representing our second contribution to the field, can be used to access the ability of the pipeline to extrapolate the knowledge extracted from the train dataset to previously unseen materials. This is an ambitious but realistic application of ML methods – it is highly desirable to develop approaches to predict entire substitution series not yet studied experimentally, as opposed to predicting single compositions close to already measured materials. In the case of TM carbides and nitrides, the pipeline produces mixed results, with some substitution series predicted extremely well, while for others the predictions are clearly not very useful. We split the series in three groups according to the ability of the trained models to predict their critical temperature, and offer some potential explanations for this divergence of the predictive power of the ML methods. As the third significant extension to the existing methods, we propose an anomaly detection algorithm as a way to estimate the accuracy of the predictions, based on the degree of proximity of the new series to the materials in the training set. This algorithm provides a tool to estimate the expected performance of trained ML models on compounds from less explored regions of materials space.

2. Method

2.1. Data sources

TM nitrides, carbides, and carbonitrides typically form interstitial compounds: the relatively large transition metal atoms form a simple lattice, with the smaller carbon and nitrogen atoms occupying the interstitial positions. The compounds that contain group IVb-Vb-VIb transition metals (Ti, Zr, Hf, V, Nb, Ta, Mo, W) exhibit similar structural, electronic, and vibrational properties. Their most common crystal structure is the cubic rocksalt (NaCl, B1-structure, Fm3 m symmetry) type, with the transition metal forming an FCC lattice and the carbon/nitrogen atoms at the octahedral interstitial positions. Despite the structural similarities, the physical properties of these materials vary significantly. In particular, their superconducting properties can be modified by substituting the transition metal element or varying the ratio of C to N, offering a large degree of tunability.

The dataset used in this paper combines information from two distinct sources. The first one is SuperCon [34] – the largest publicly available database of superconductors. SuperCon compiles records of superconducting properties of materials collected from published literature. It is maintained by the Japanese National Institute for Materials Science. Several experimental properties can be accessed through this database for a wide range of different classes of superconductors, including intermetallic, oxide, and organic materials. We extracted 39 transition-metal carbides and nitrides with NaCl-type structure.

Table 1
Transition metal carbide and nitride series in the dataset, together with their respective T_c ranges.

Starting composition	Substitution	Minimum T_c (K)	Maximum T_c (K)
HfC	MoC	3.4	9.0
HfC	NbC	4.5	7.0
TaC	MoC	7.5	8.9
TaC	NbC	9.4	11.1
MoC	NbC	12.0	13.4
NbN	HfN	5.5	14.1
NbN	ZrN	10.0	14.2
NbN	NbC	12.4	17.9
NbN	TiN	6.6	17.1
NbN	HfC	12.8	15.5
NbN	TaC	11.6	16.5
NbN	TiC	14.9	18.0
NbN	VC	5.0	12.2
NbN	ZrC	14.3	16.3
NbN	VN	3.1	11.9
VN	VC	2.1	9.8
VN	NbC	2.1	9.2
VN	TaC	2.5	7.4
VN	TiC	1.9	8.6

To supplement the data from Supercon, we extracted another 108 compounds from scientific literature [11,13]. These works studied pseudo-binary, pseudo-ternary, and pseudo-tertiary compounds and experimentally determined their critical temperatures. Most of these are based on the combination of NbN and VN with nitrides or carbides of other elements. Since most of the data in these publications is contained in graphs, we digitized many of the figures (using WebPlotDigitizer [35]) to extract the experimental values. The list of all substitution series are given in Table 1. See also Fig. 1, which shows some example series demonstrating one of the most intriguing aspects of the superconductivity in these compounds – the highly non-monotonic dependence of T_c on substitution.

It is important to point out that experimentally determining T_c poses inherent uncertainties. For the information extracted from scientific literature, experimental uncertainties were often provided in the plots, and generally fall between 0.2 K and 1 K. However, in some specific cases, the T_c values in the dataset are merely rough estimates. For example, it has been reported that the T_c of TiC falls somewhere between 1.1 K and 1.8 K, so a value of 1.5 K was used [36]. In the cases of HfC and VC, where superconductivity has not been observed experimentally above a certain threshold (1 K for HfC [19]), a T_c of 1 K was used. Additional uncertainty in some T_c values is introduced by the unavoidable errors associated with digitizing plots (which are sometimes with rather low resolution). These we estimate to be in the range of 0.3–0.5 K.

Despite these potential issues, the dataset described here is significant as the first – to the best of our knowledge – effort to compile the information available on superconducting TM nitrides, carbides, and carbonitrides.

2.2. Featurization

The dataset described above contains only the compounds' chemical formula, encoding the elements contained and their stoichiometry, as well as the measured critical temperature T_c values of these compounds. To proceed, we need to associate numerical predictors (also called features) based on the chemical composition, which we can then use to train an ML model for predicting a target quantity. The target in our case is T_c.

Several types of predictors can be generated, depending on the properties of interest. For example, these can incorporate geometrical or structural properties, which have been known to play an important role in superconductivity. In the dataset under study all compounds have the same NaCl-type structure and a single structural feature – the lattice parameter – allows us to describe the difference between materials. Lattice parameters can be experimentally determined or calculated theoretically using *ab initio* tools such as Density Functional Theory (DFT). In this study, the experimental values, where given, were used. Since for many materials in the dataset no experimental values were available, to impute them we assumed the materials follow the Vegard's law, i.e. the lattice parameters linearly depend on the percentage of elementary compounds in the alloys (which is a good approximation to the behavior observed experimentally – see, for example, Fig. 1 in Ref. [12]). For example, if an alloy consists of 30% of NbN and 70% of VN, then the lattice parameter of the alloy will be equal to the sum of 30% of the lattice parameter of NbN and 70% of that of VN. The simple heuristic of the Vegard's law allows us to (approximately) reconstruct the structure of all compounds in the dataset using only the relatively few measured lattice parameters. (It has to be noted that this simple procedure carries the risk of introducing systematic bias in the ML models.)

Once the structure is determined, the focus shifts towards elemental and electronic attributes. To generate these, we used the Materials Agnostic Platform for Informatics and Exploration, also known as Magpie [37], conveniently provided by the *matminer* package [38]. Magpie represents a framework in which a set of features based on the elemental composition of a material is calculated. The set consists of statistics (such as weighted mean, standard deviations, minimum, maximum, and range) of 22 elemental properties, including atomic weight, electronegativity, the electron numbers and fractions of electrons in different valence shells.

At the end of the featurization step, the sole structural feature (the lattice parameter) is combined with 132 Magpie predictors, encoding elemental and electronic properties, creating a total feature set of 133.

2.3. Feature selection and dimensionality reduction

So far, the dataset created consists of 147 rows of compounds with 135 columns, 133 of which (apart from the non-numerical formula and the target critical temperature) can be used as features for prediction. Using a large number of features for such a small number of samples can easily lead to overfitting or prevent the model from learning by introducing high levels of noise (by including many irrelevant features). Thus, the ultimate number of features used in the ML modeling should be as low as possible. To achieve this we study the importance of each feature and carefully select the materials descriptors used to build a model. We also apply dimensionality reduction techniques for easier visualization of the data, which can provide insights by showing important trends.

2.3.1. Feature selection

The selection of relevant features is very important when dealing with highly dimensional data. In this work, we focus on three feature selection methods, based, respectively, on using a variance threshold, the mutual information score and the Random Forest feature importance score. The variance threshold selection is a simple baseline approach that removes all features whose variance does not meet a certain threshold value selected in advance. This method is based on the assumption that features with little variance only have a small effect, if any, on the values of the target value to predict. On the other hand, features with higher variance values may contribute more to the target value. However, this approach remains fairly simplistic, as it is only based on a single threshold value, which is generally difficult to select in advance. Therefore, in practice, this method is used almost exclusively to remove the features with zero variance, which certainly do not contribute to the target prediction. Feature selection is also performed with the mutual information score, which is calculated with respect to two variables, generally one feature and the target value. This score measures the reduction of uncertainty in the target variable given a value of a feature variable (see Ref. [39]). Therefore, features with a higher mutual information score are considered to contribute more to the target value, and vice versa. Finally, a useful method to consider when training with tree-based ML models (for example Random Forest, explained below) is their calculated feature importance score, which can provide an idea about any important features that might have been missed from the previous two methods.

2.3.2. Dimensionality reduction

Principal Component Analysis (PCA) [40] is one of the most widely used dimensionality reduction techniques. PCA converts an initial dataset with dimension n to a dataset with a lower dimension k, while maintaining as much information as possible from the original dataset. This is done by maximizing the variance of the projected datapoints in the new basis set. t-distributed Stochastic Neighbor Embedding [41], or t-SNE, is also frequently used to reduce highly dimensional data. This approach is based on preserving local similarities between different features rather than over the whole dataset. In this work, we focus on a novel technique called Uniform Manifold Approximation and Projection (UMAP) proposed by McInnes et al. [42]. This technique makes use of Riemannian geometry and algebraic topology for reduction of highly-dimensional data. It combines the visualization ability of t-SNE while preserving more of the global structure of the data, similar to PCA.

2.4. Machine learning models

ML is a very promising approach for predicting materials properties. The ML models are trained on existing data, effectively learning meaningful patterns which may be too complex to be easily discovered and utilized by human researchers. If these patterns are transferable to some new, previously unseen materials, the trained model can be used to accurately predict the property of interest, for example the transition temperature of a superconductor.

Many different ML approaches have been used to predict various materials properties, including T_c. In this paper, we make use of the Random Forest [43] regression algorithm, one of the models that has proven successful when dealing with complex patterns and a relatively high number of input features. This algorithm is an ensemble method combining multiple decision trees based on the input features. Once all the decision trees have been fitted on the training data, the mean of the predictions of all the decision trees are calculated to give the final output of the model. By calculating the standard deviation of the values generated by each tree, a Random Forest regressor is also capable of providing an estimate of the prediction uncertainty. Furthermore, impurity-based feature importances can be easily calculated from the trained model, which can be used to supplement feature selection methods detailed in the previous section.

Powerful methods such as Random Forest can perfectly reproduce the training data, effectively "learning" not only the meaningful patterns present in the data but also the "noise" particular to a specific training dataset. Thus, it is crucial to validate the ML model on a test data which has not been used during the training stage. Typically this test data is extracted from the full available dataset in a random fashion. In some cases, however, this does not represent a true test of the model. In particular, when applying a ML model for predicting T_c, a realistic target could be an entire new substitution series (for example, adding VN to TiC), without any points of the series present in the training dataset (except potentially the end points). Thus, to provide a realistic estimate of the model in this likely user case, we have to split the dataset appropriately, and train and test its performance on different series. Notice that this is a significantly more difficult task

Table 2
The selected 12 features (previously defined lattice parameter as well as elemental features based on the Magpie framework [37]) along with their mutual information (MI) score with T_c. The top 10 were selected based on this score, and the last two were found using feature importance values from the Random Forest algorithm.

Feature name	Description	MI score
Range MendeleevNumber	Range of elemental Mendeleev numbers	0.64
Avg dev CovalentRadius	Average deviation of elemental covalent radii	0.59
Lattice parameter	Lattice parameter	0.58
Mean GSvolume pa	Mean DFT-computed volume of elemental solid	0.56
Range Electronegativity	Range of elemental electronegativities	0.55
Mean CovalentRadius	Mean elemental covalent radius	0.50
Avg dev Number	Average deviation of elemental atomic numbers	0.48
Avg dev Row	Average deviation of elemental periodic table rows	0.46
Mean Number	Mean elemental atomic number	0.46
Range CovalentRadius	Range of elemental covalent radii	0.45
Mean Ns Valence	Mean number of filled s orbitals	0.32
Mode Ns Valence	Mode number of filled s orbitals	0.25

than predicting the T_c of randomly selected materials, since using an entire series as a test set almost certainly precludes close proximity between the points in the training and test sets. On another hand, this forces the model to learn patterns which can be generalized from series to series, as opposed to just memorizing nearby data points.

3. Results

Fig. 1 represents some of the substitution series present in the dataset, collected as described in the Method section. The end goal of the ML modeling is to be able to predict the evolution of the critical temperatures T_c of the materials in the dataset as a function of their composition. We start by selecting the most important features correlated with T_c. We also visualize the originally high-dimensional data to better understand its structure. We then create a pipeline for predicting the critical temperature by training a Random Forest regression model on different test sets. An exhaustive search of the top performing models is performed to select the optimal combination of input features that leads to the best model metrics. Using these features, we train a Random Forest model to predict both randomly selected points and entire substitution series. The latter tests the ability of the ML approach to model previously unexplored materials combinations – an important practical application. All of the above has been conducted with Python, specifically using software packages *umap-learn* [42] for dimensionality reduction and *scikit-learn* [44] for feature selection and ML model training.

3.1. Feature selection

3.1.1. Variance threshold

As outlined in the Method section, the variance threshold is useful for removing the features with low variance in values, based on the assumption that these would not contribute to the final model. In this work, we set the variance threshold to be 0, in order to remove all the features that certainly do not provide any information. Of the initial 133 features, it was found that 19 have zero variance and were thus dropped from the dataset.

3.1.2. Mutual information score

In order to choose the most relevant features for predicting the target, the mutual information score between each input feature and T_c was calculated. This was done for the 114 features left in the dataset after removing the zero-variance features. These were then ranked from the highest mutual information score, and the top 10 were selected (shown in Table 2).

3.1.3. Feature importance from random forest

Our last method for feature selection is the feature importance score available when training a Random Forest algorithm. For that purpose, a Random Forest regressor is trained on a random train subset of the data, and the relative feature importances are calculated. The features with the highest importance that are not detected using the mutual information score are then selected. This is the case of the magpie features *mean Ns valence* and *mode Ns valence*, which represent respectively the mean and mode values of the electron valence in the s orbitals.

At the end of the feature selection procedure, we are left with a total of 12 features: the top ten features according to their mutual information score, and an extra two features selected from the Random Forest feature importance. The full list is shown in Table 2. These are used later on for visualization using dimensionality reduction and for the exhaustive search in order to select the best feature combination for the ML models.

3.2. Dimensionality reduction

To visualize the dataset, we project the materials points from the space of the previously selected 12 features onto a two-dimensional space using the dimensionality reduction technique UMAP. This allows us to examine the structure of the data before training the ML models on it. The UMAP results are shown in Fig. 2**b**. As can be seen there, the data points tend to separate into distinct clusters. Interestingly, these clusters cut across both the TMs' group and period in the periodic table. Most Hf- and Zr-based series belong to the upper left cluster, whereas a majority of V and Ti series tend to cluster towards the lower right of the UMAP visualization (with a few exceptions). The series $(VC)_x(VN)_{1-x}$ and $(MoC)_x(NbC)_{1-x}$ appear to form two separate smaller clusters. The clusters show a relatively smooth distribution in terms of T_c values.

3.3. Machine learning models

3.3.1. Prediction of T_c for random test set

As a first test of the ML approach, we develop the pipeline by training and testing it on a random subset of the original data. Benchmarking the model on a test set containing randomly selected materials is appropriate if it will be mostly used to predict T_c of individual novel instances, some of which might be quite close to already measured materials. An exhaustive search is performed in the feature space to select the best performing ML model. Different combinations of the 12 features discussed in the previous section are generated, and for each combination an ML model is trained using a five-fold cross-validation repeated three times, with the data randomly shuffled every time. The metrics used for evaluation of the performance of the model outside of its training data are the mean absolute error (MAE), the root mean squared error (RMSE) and the R^2 score. Each of these metrics has their own advantages and shortcomings, which explains the importance of

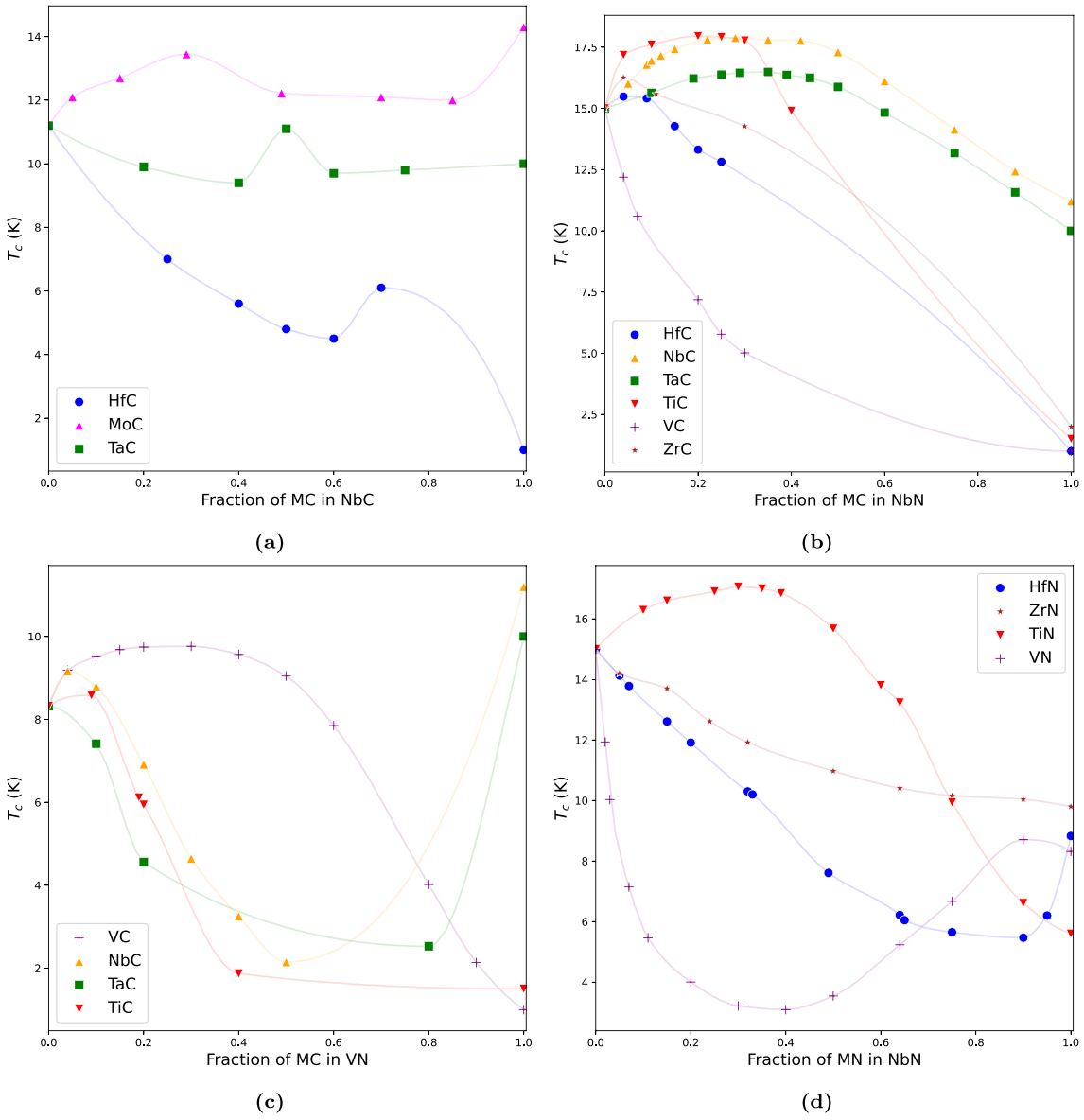

Fig. 1. The experimentally observed evolution of T_c of some series present in the dataset in terms of the fraction of one of the compounds. M denotes a given transition metal. **(a)**: metal carbides MC in NbC. **(b)**: metal carbides MC in NbN. **(c)**: metal carbides MC in VN. **(d)**: metal nitrides MN in NbN. The lines connecting the measured points are Piecewise Cubic Hermite Interpolating Polynomial fits (PCHIP), provided only as a visual guide. Notice that the data points are subject to experimental uncertainties and digitalization errors, as discussed in the text.

using all of them to select the optimal model. While MAE measures the average magnitude of error and gives equal weight to individual differences, RMSE and R^2 are both sensitive to outliers and penalize large individual errors. In this example, priority is given to RMSE and R^2 in selecting the optimal model features, in order to avoid large prediction errors as much as possible.

The best model selected from the exhaustive search has an R^2 score of 0.82, and RMSE/MAE values of 1.94 K and 1.37 K respectively. It is based on the following four features: the mean *Ns valence*, the mean *covalent radius*, and the ranges of the *Mendeleev Number* and the *electronegativity*. (Note that the second one has been already flagged as important in a previous ML study of low-temperature conventional superconductors [27].) In order to further examine the performance of this specific model, we also perform a Leave-One-Out Cross-Validation (LOO-CV). This method repeatedly splits the dataset into a testing set consisting of a single data point, and a training set consisting of all remaining instances. It helps provide a less noisy estimate of the model's error compared to using a conventional random train-test set split approach. When trained using LOO-CV, an R^2 value of 0.85 is obtained, as well as values of 1.82/1.26 K for RMSE and MAE respectively. Fig. 2**a** visualizes the obtained results using LOO-CV.

3.3.2. *Predictions within existing substitution series*

Using the optimal features selected in Section 3.3.1 and the entire dataset, we can develop an ML model to predict new points. One interesting application of this model directly stems from the existing data. While some substitution series have been studied experimentally quite extensively, others have been measured only at a few compositions and are therefore very sparsely represented in the dataset. We can use the ML model to predict the critical temperature of new instances within existing substitution series. This task can make use of nearby data points in the training data, without having to generate predictions for entirely new series. Fig. 3 summarizes the results obtained for compositions within a few substitution series. There we show both the predicted critical temperature, as well as the prediction uncertainty of the model. (The latter is estimated as the standard deviation of the predictions of the ensemble of trees forming the Random Forest model.)

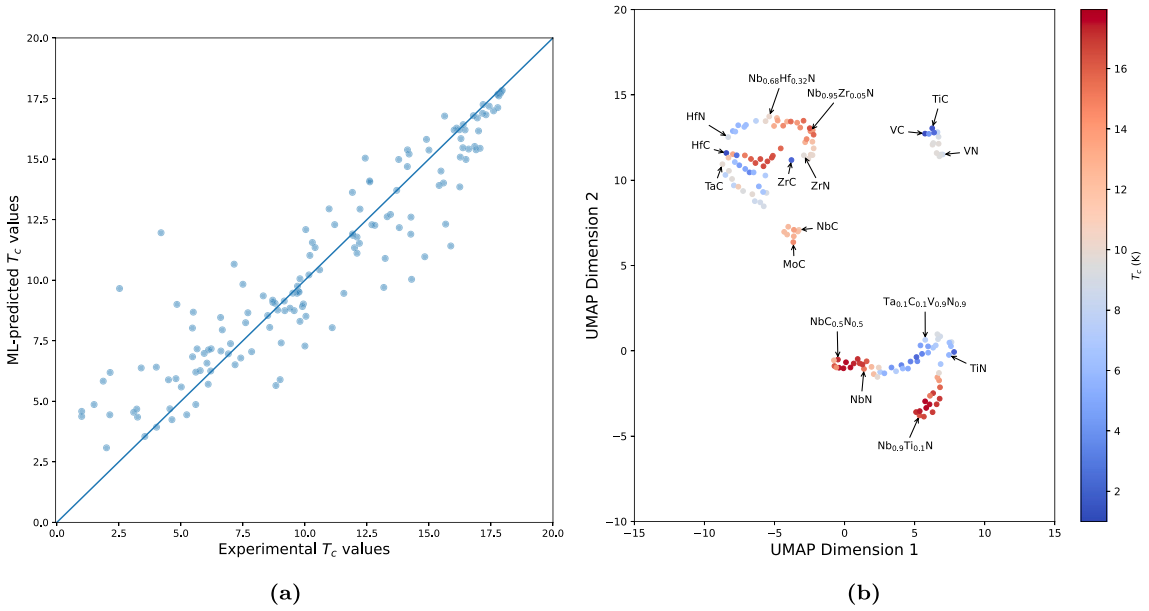

Fig. 2. **(a)**: Visualization of the ML-predicted vs experimental T_c values of the models trained via Leave-One-Out cross-validation on the optimal feature combination in the baseline model. **(b)**: Visualization of the 12-dimensional feature space (the features are listed in Table 2) using the UMAP technique. The 12-dimensional space is projected onto two axes, plotted above. The points representing different materials in the dataset are colored according to their T_c value.

As we can see, the predictions are broadly in line with the behavior observed in the other series. The trained model clearly was able to learn one of the most interesting features of the TM carbide/nitride superconductivity, namely the non-monotonic substitution dependence of T_c. In the cases of $(VN)_x(NbC)_{1-x}$ and $(VN)_x(TiC)_{1-x}$ series, the model predicts potential maxima of T_c in regions which have not been explored so far. It would be very interesting to test these predictions experimentally.

3.3.3. Prediction of T_c for entire substitution series

After developing and benchmarking ML models using a conventional approach based on random splits of the dataset, we shift our attentions to the task of predicting the behavior of entirely new and previously unexplored substitution series. As already discussed in the introduction, this is a much more challenging problem. It is also of significant practical interest; exploring all possible elemental combinations experimentally would be very time- and resource-intensive and an ML model with good predictive power could allow researchers to focus on the most promising substitution series.

To optimize the models for this particular task, we again perform an exhaustive search for the best feature combination. The same feature combinations discussed in the previous section are used. This time, for each combination the data is split into a testing set consisting of one of the doping series, and a training set of all remaining series. Therefore, the number of Random Forest models trained for each feature combination corresponds to the number of series present (19 in our case). Once the metrics are calculated for each model with each series as a test set, these are averaged to assess the overall performance for a given combination of features.

The selected best model in this case is based on two features: the *mean Ns valence* and the *range of the electronegativity*. It has an MAE average of 1.88 K and an RMSE of 2.16 K, both of which remain comparable to that of the baseline model (MAE of 1.37 K and RMSE of 1.94 K). However, the average R^2 score of the selected model is −3.3 compared to the previous value of 0.82, which is due to the wide distribution of R^2 scores in the doping series' predictions. Indeed, the model performs much better on some series (with R^2 scores higher than 0.7) compared to others (with very low or even negative R^2 scores). This shows the difficulty of predicting the critical temperature of an entirely new substitution series with previously unexplored materials.

Table 3
Statistics of the obtained metrics for the selected Random Forest models based on the *mean Ns valence* and *range electronegativity* features, evaluated on all substitution series in the dataset as testing sets.

	R2 score	MAE	RMSE
minimum	−26	0.98	1.03
25th percentile	−2.3	1.18	1.47
mean	−3.3	1.88	2.16
median	−0.33	1.96	2.32
75th percentile	0.034	2.52	2.7
maximum	0.79	2.95	3.49

Table 3 shows statistics of the metrics of the selected model over the 19 substitution series.

In Fig. 4**a** we show the comparison between the predicted and experimental critical temperature values for the different substitution series. As can be seen there, some series are tightly clustered around the diagonal [for example, $(HfC)_x(NbN)_{1-x}$ and $(VC)_x(NbN)_{1-x}$]. Other series show more dispersion, but they also tend to follow the diagonal [e.g. $(HfN)_x(NbN)_{1-x}$]; a few series show no correlation between measured and predicted values – examples are $(NbC)_x(NbN)_{1-x}$ and $(NbN)_x(VN)_{1-x}$. Based on this observation, and the statistics of the performance metrics for the series, we define three groups of substitution series. These separate the series for which the model has been deemed successful at predicting critical temperature values from the ones where the predictions are less accurate. We select these groups based on Fig. 4**a**, and confirm our choice using MAE/RMSE values (both of them following similar trends) and R^2 scores. The top performing series tend to have an MAE score lower than the median MAE score of 1.96 K and an R^2 score higher than 0 (i.e., the model is doing better than using the average of the training set). These we combine in Group 1. In Group 3 we collect the worst predicted substitution series, with MAE generally higher than the 75th percentile value of 2.52 K (this often coincides with negative R^2 scores). In Group 2 we put all the series which do not belong to either Group 1 or Group 3. There are six series in Group 1, five in Group 2, and eight in Group 3 (see Table 4 for details), respectively with 31, 45 and 57 data points, corresponding to individual material instances.

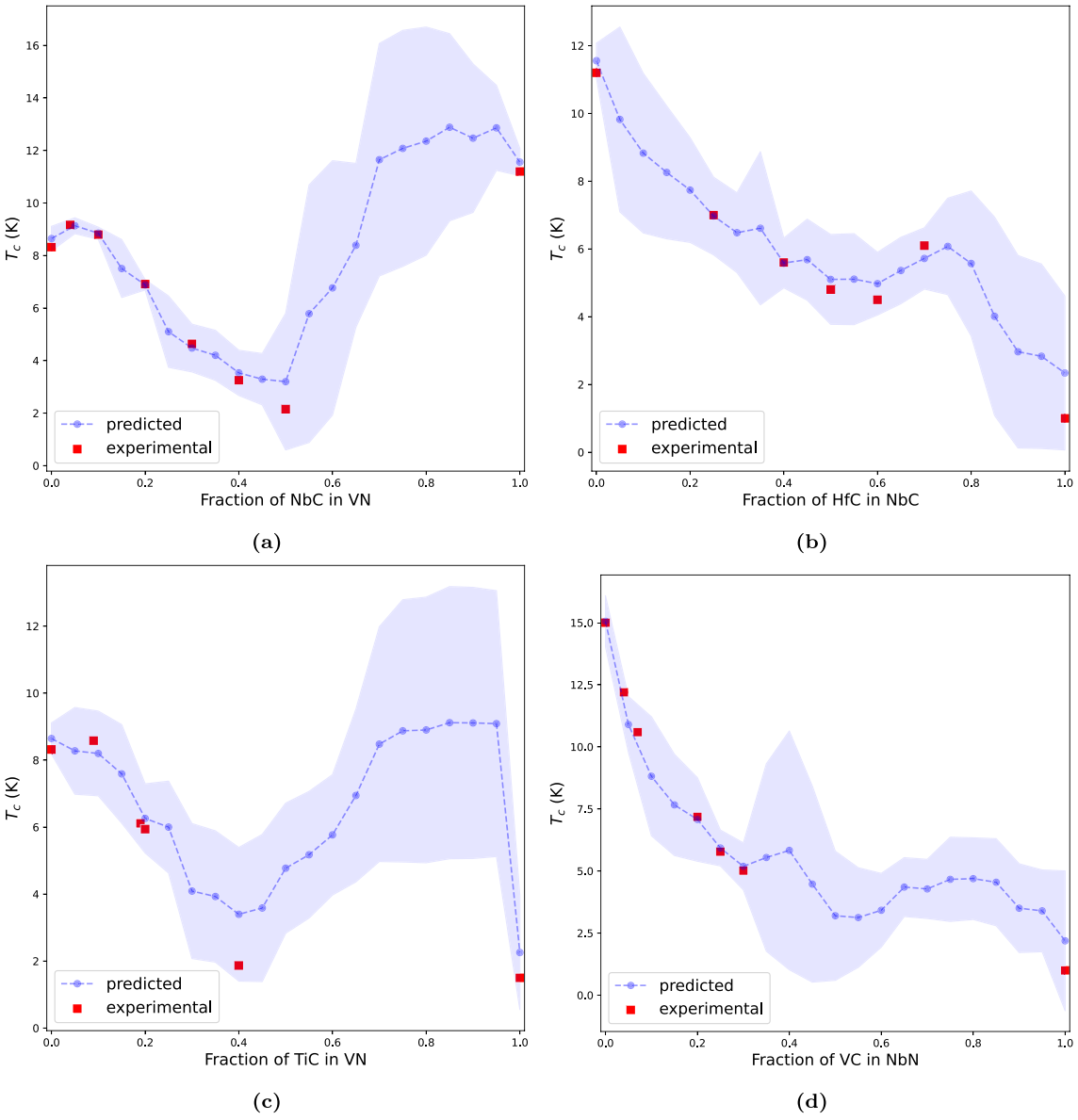

Fig. 3. The ML modeling of four substitution series, showing the predicted T_c values of several compositions not studied experimentally so far, along with the measured points. **(a)**: NbC in VN. **(b)**: HfC in NbC. **(c)**: TiC in VN. **(d)**: VC in NbN.

In Fig. 4**b**, we show the same groups but using the UMAP projection of the feature space. Two of the series forming their own small clusters $((VC)_x(VN)_{1-x}$ and $(MoC)_x(NbC)_{1-x})$ belong to Group 3, which can be explained by the absence of nearby points allowing the model to successfully extrapolate. The upper left cluster mostly showed good performance (Groups 1 and 2). This region is densely populated, which has certainly helped the model to accurately predict the behavior of new series using neighboring ones. One can note a few exceptions $[(TaC)_x(NbC)_{1-x}$ and $(TaC_x(MoC)_{1-x}$ for example], which tend to be at the borders of the cluster. This effect is less clear for the lower right cluster, which shows varying performance levels.

A more clear distinction can be drawn between the distribution of feature values within each group of series and the distribution of the entire dataset (see Fig. 5**a**). For the *range Electronegativity* feature, the distribution of the well-predicted series (Group 1) is shifted higher up and is narrower, falling generally between values of 1.2 and 1.5. This effect is also clear for the *mean Ns Valence* feature (between 1.6 and 1.75), although the distribution of the less well-predicted series (Group 3) tends to center around slightly higher valence values (above 1.65) than that of the Group 1 series, which seems to have a slightly lower median value (below 1.65). In both cases, the interval of feature values of the Group 1 series tend to be more narrow and centered around a smaller range, while Group 3 series feature values tend to be quite sparse. This suggests that the model has good predictive power in specific feature value intervals, while showing more fragile performance outside of them.

To quantify these observations and develop a more rigorous way to distinguish the series belonging to the different groups, we make use of an anomaly detection algorithm. Indeed, ML models generally tend to perform better on test data similar to the training set and, conversely, have a much higher chance of performing poorly on data points from regions outside of the training data. Thus, in an anomaly detection setting, it could be expected that an ML model would better predict normal points (also known as inliers), and show inferior performance for anomalous points (outliers).

We test this hypothesis for the different substitution series using Isolation Forest as the anomaly detection algorithm [45]. This method aims to isolate each of the points in the dataset using different decision tree splits. It then calculates a decision function score (IF score) based on the number of splits, and uses that score to distinguish inliers from

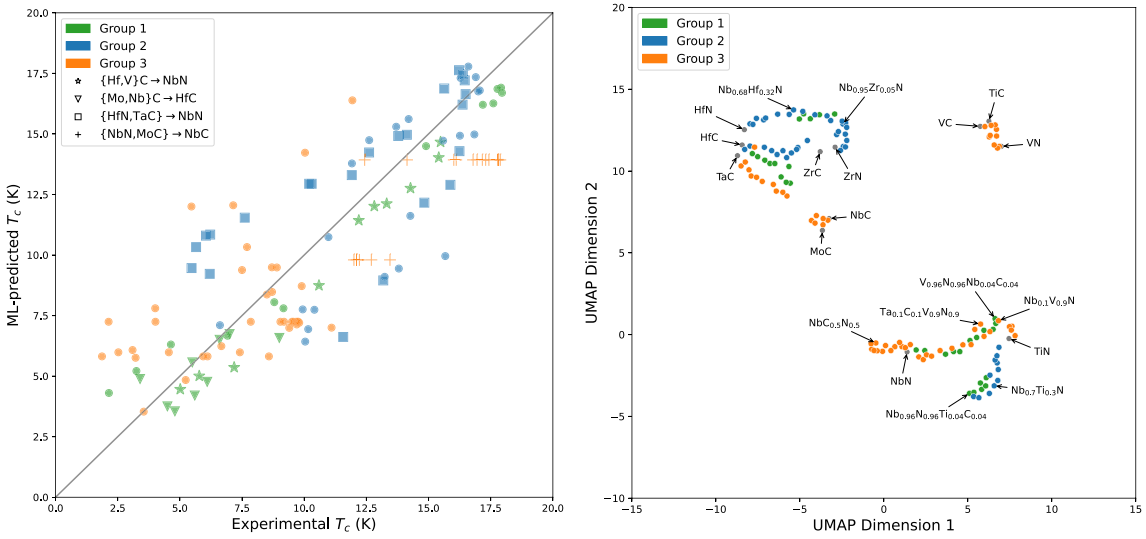

Fig. 4. The series divided by groups (see Table 4). **(a)**: Predicted versus experimental T_c values; **(b)**: UMAP projection of the 12-dimensional feature space (cf. Fig. 2**b**), with points colored by group.

Table 4
Classification of the doping series into the previously defined groups. Group 1 consists of the series that show the best performance, and Group 3 of those with the worst. Group 2 consists of series that are neither in Group 1 nor 3. x denotes the fraction of one compound in the resulting alloy mixture, and its value falls between 0 and 1. The table also shows the mean and minimum of the calculated decision function value of the Isolation Forest (IF) anomaly detection algorithm (IF score).

Series	Group	Mean IF score	Minimum IF score
$(VC)_x(NbN)_{1-x}$	1	0.064	0.030
$(NbN)_x(TiC)_{1-x}$	1	0.029	0.012
$(HfC)_x(NbN)_{1-x}$	1	0.020	0.002
$(VN)_x(NbC)_{1-x}$	1	−0.007	−0.046
$(HfC)_x(MoC)_{1-x}$	1	−0.011	−0.021
$(HfC)_x(NbC)_{1-x}$	1	−0.037	−0.067
$(ZrC)_x(NbN)_{1-x}$	2	0.022	−0.009
$(NbN)_x(TiN)_{1-x}$	2	−0.012	−0.059
$(NbN)_x(TaC)_{1-x}$	2	−0.019	−0.081
$(NbN)_x(ZrN)_{1-x}$	2	−0.040	−0.108
$(HfN)_x(NbN)_{1-x}$	2	−0.048	−0.152
$(NbC)_x(NbN)_{1-x}$	3	0.041	0.015
$(NbN)_x(VN)_{1-x}$	3	0.026	−0.015
$(VN)_x(TaC)_{1-x}$	3	−0.016	−0.044
$(NbC)_x(TaC)_{1-x}$	3	−0.033	−0.065
$(MoC)_x(NbC)_{1-x}$	3	−0.041	−0.057
$(VN)_x(TiC)_{1-x}$	3	−0.047	−0.052
$(MoC)_x(TaC)_{1-x}$	3	−0.056	−0.089
$(VN)_x(VC)_{1-x}$	3	−0.092	−0.122

outliers. The IF score is based on the anomaly score $s(x, n)$ defined in [45], and is calculated using the following formula:

$$\text{IF score} = -s(x, n) + 0.5 \tag{1}$$

0.5 is a standard offset [44]. For a given point, the more splits required to isolate it, the higher the IF score, and therefore the more likely it is to be an inlier and thus belong to the main dataset distribution. On the other hand, if only a few splits are required to isolate a certain point, then the IF score will be lower, making it an outlier. We apply the algorithm to our dataset using the substitution-series-based train-test split and the 12 features described in Section 3.2 (see Table 2). For each doping series, the data is split into a testing set consisting of that specific series, and a training set of all remaining series. An Isolation Forest algorithm is fitted on each training set and the IF scores are then computed for each material in the testing set. A mean IF score is then calculated for all materials in the series.

Fig. 5**b** shows the distribution of the obtained mean decision function values for each series, colored by the group to which it belongs (Table 4).

It is observed that the distribution of the mean decision function score is centered around higher values for Group 1 series than Group 2 series, whose score is in turn centered around higher values than Group 3 series. Therefore, materials in Group 1 series tend to be inliers, while Group 3 materials are more likely to outliers. Exceptions are $(NbC)_x(NbN)_{1-x}$ and $(NbN)_x(VN)_{1-x}$ series, which belong to Group 3, despite having large decision function scores. In order to understand these exceptions, a closer look at the performance of the model on these two series is required.

In the case of $(NbC)_x(NbN)_{1-x}$, the UMAP projection (Figs. 2(b) and 4(b)) shows that the materials of this series are part of a cluster (in the lower-right corner of the projection), which explains why the IF scores are relatively high. However, they are located at the edge of this cluster and away from the main distribution of materials in the region, which probably explains the poor performance of the ML model. This is reflected in almost constant T_c values predictions within that series (Fig. 6(f)), suggesting the points are not sufficiently separated in the feature space.

In contrast, $(NbN)_x(VN)_{1-x}$ series is located well within the lower-right cluster of the UMAP projection (Figs. 2(b) and 4(b)), leading to high IF scores. This suggests the ML approach should be able to accurately model this series. Indeed, looking at the ML predictions in Fig. 6(e), we can see the model is able to reproduce the general trends of its evolution rather well. However, the selection criteria mentioned earlier puts it in Group 3 (Table 4). Thus, R^2 and RMSE should only be used as rough guidelines when developing predictions for entire series, in which case uncovering the general trend is arguably at least as important as predicting the exact T_c values.

These results demonstrate that anomaly detection algorithms can provide a general guidance on the expected performance of the ML algorithm on previously unseen substitution series. If the studied series has a high mean decision function score, it consists mostly of inliers belonging to a similar distribution as the training set, which would likely result in good model performance. On the other hand, if the doping series consists mostly of outliers compared to the training dataset (low decision function scores), the ML model will probably show an inferior performance, similar to the one on Group 3 series.

Fig. 6 compares the experimentally measured and model-predicted evolution of the critical temperature of the doping series in terms of

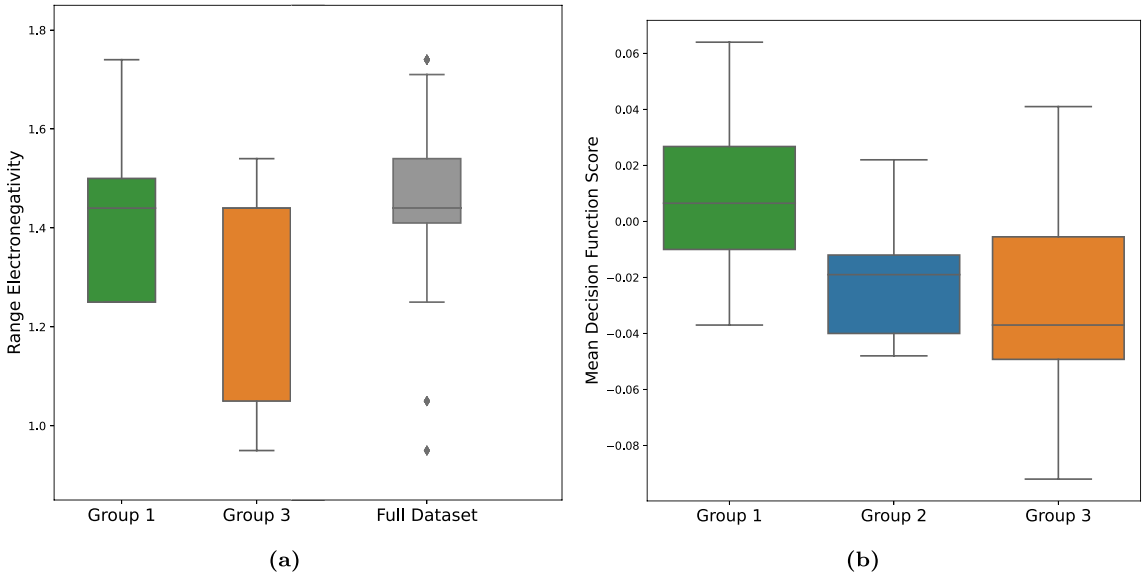

Fig. 5. **(a)**: Visualization of the distribution of the feature *range Electronegativity* for the best and worst performing series as well as the overall distribution for the selected best model. **(b)**: Visualization of distribution of the obtained mean decision function values for each series, colored by the group to which it belongs.

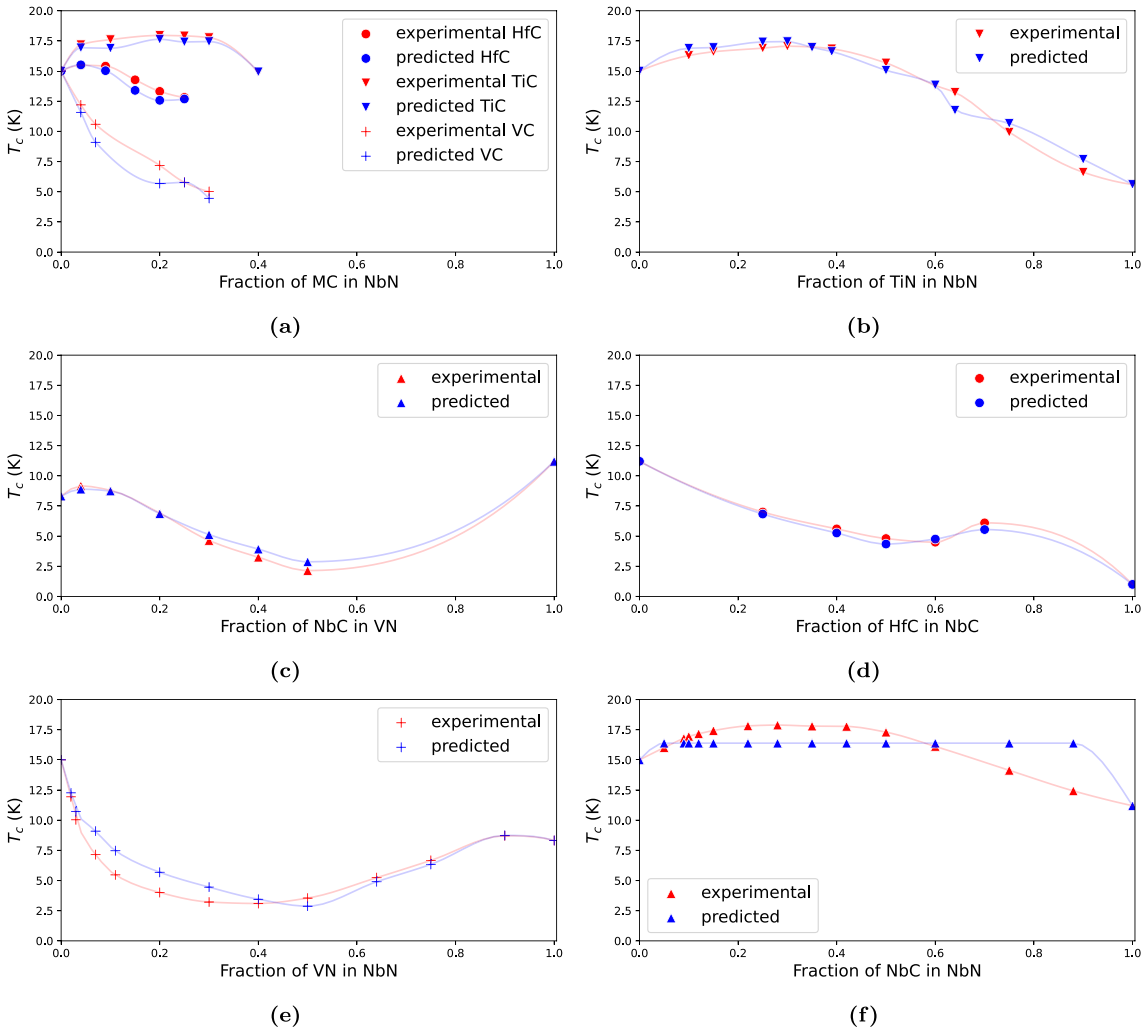

Fig. 6. Comparison of the experimental (red) and the ML-predicted (blue) evolution of the critical temperature T_c of doping series in terms of fraction of compounds present. **(a)**: MC in NbN. **(b)**: TiN in NbN. **(c)**: NbC in VN. **(d)**: HfC in NbC. **(e)**: VN in NbN. **(f)**: NbC in NbN.

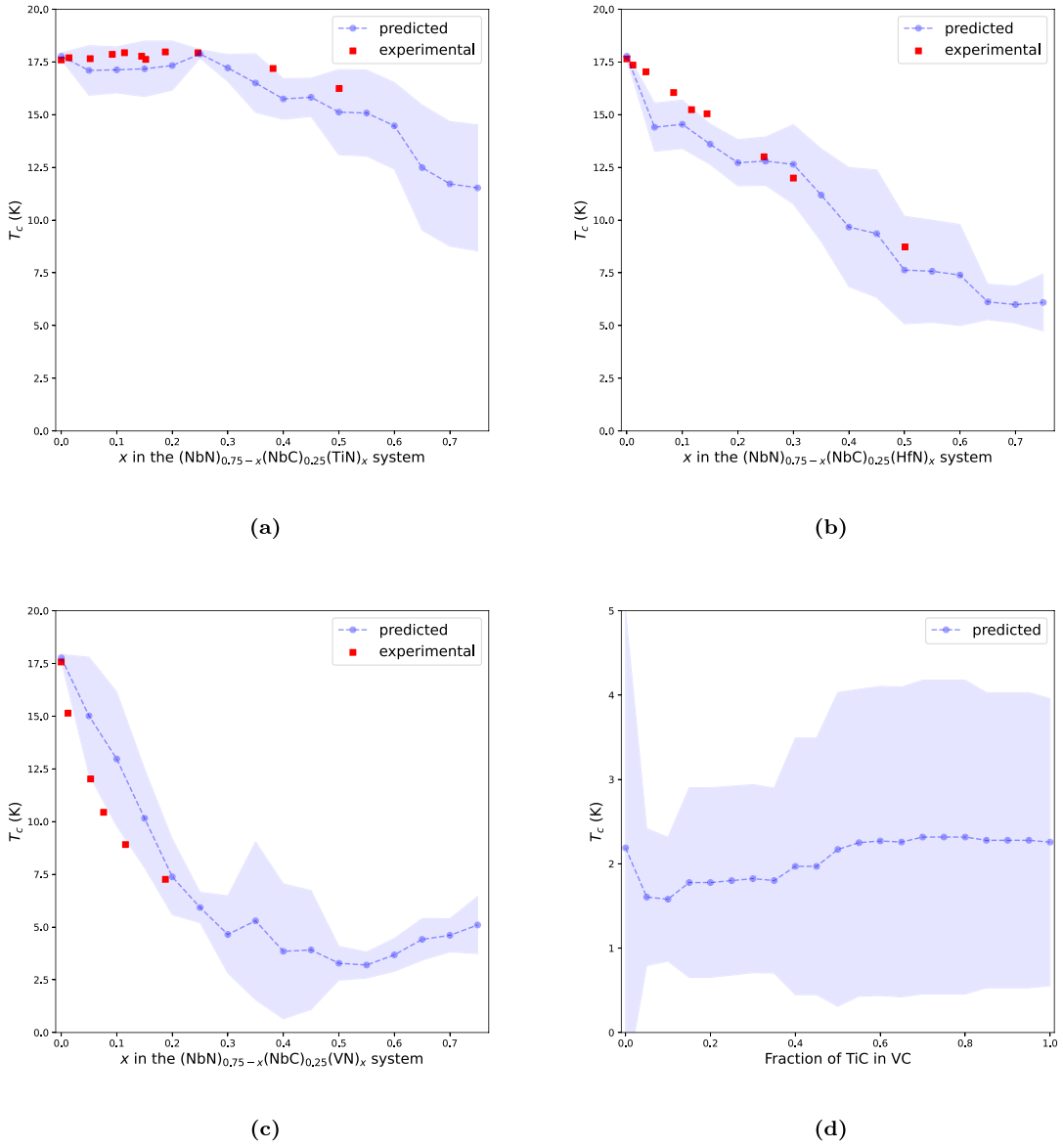

Fig. 7. Plots of predicted vs experimental T_c for pseudo-ternary compounds of type $(NbC)_{0.25}(NbN)_{0.75-x}(MN)_x$ or $(NbC)_{0.25-x}(NbN)_{0.75}(MC)_x$, with M a group IV or V transition metal, as a function of x, the fraction of MN or MC added. **(a)**: the $(NbC)_{0.25}(NbN)_{0.75-x}(TiN)_x$ system. **(b)**: the $(NbC)_{0.25}(NbN)_{0.75-x}(HfN)_x$ system. **(c)**: the $(NbC)_{0.25}(NbN)_{0.75-x}(VN)_x$ system. **(d)**: the $(TiC)_x(VC)_{1-x}$ system, experimentally found to be non superconducting until at least 2 K.

the composition. It is observed that for the majority of the substitution series represented, the predicted critical temperatures follow a similar trend to the experimental ones (subfigures **(a)-(e)** in Fig. 6). However, for some of the series, the predicted evolution is not comparable to the experimental ones. In the specific example of subfigure **(f)** of Fig. 6, the model consistently predicts the same T_c for all the fractions. This is due to the projection onto the 2D feature space used, in which all the resulting alloys of that substitution series have the same feature values. This prevents the model from distinguishing different fractions in that series.

3.3.4. *Prediction of T_c for pseudo-ternary compounds*

In an effort to increase the critical temperature of the pseudo-binary system $(NbC)_{0.25}(NbN)_{0.75}$ (T_c of about 17.8 K), Pessall et al. [13] conducted experiments involving alloying additional Group IV and V transition metals, thus leading to pseudo-ternary compounds of the form $(NbC)_{0.25}(NbN)_{0.75-x}(MN)_x$ or $(NbC)_{0.25-x}(NbN)_{0.75}(MC)_x$, with M being the additional transition metal. They subsequently analyzed the resulting evolution of T_c values as a function of x, the fraction of the added alloying compound MN or MC. Since our ML pipeline has been trained exclusively on pseudo-binary compounds, predicting the evolution of the critical temperature of pseudo-ternary compounds is a significant further test for the models. Fig. 7 represents the results obtained for the $(NbC)_{0.25}(NbN)_{0.75-x}(TiN)_x$, $(NbC)_{0.25}(NbN)_{0.75-x}(HfN)_x$ and $(NbC)_{0.25}(NbN)_{0.75-x}(VN)_x$ pseudo-ternary systems. As can be seen in the figure, there is a very good agreement between the predicted and the experimentally measured values of T_c. This applies both to the range of values and trends with various substitutions. Consistent with the experimental results, the addition of an alloying element was not predicted to lead to an improvement in the T_c of $(NbC)_{0.25}(NbN)_{0.75}$. In fact, the addition of a third alloying carbide or nitride was found to decrease T_c. The ML model successfully predicted that trend, including the interesting observed tendency of almost constant T_c before the added amount of TiN reaches some substantial threshold ($x \approx 0.25$) (see Fig. 7**a**).

We also tested the pipeline on $(TiC)_x(VC)_{1-x}$ series, not showing signs of superconductivity up to 2 K [13] (see Fig. 7**d**). The predictions are in agreement with this result.

4. Discussion and conclusion

Here we presented an application of ML methods for modeling and understanding the superconducting behavior of TM carbides, nitrides, and carbonitrides. As a crucial prerequisite for this work, we assembled for the first time a dataset of 147 relevant materials found in the SuperCon database and throughout the scientific literature.

Once the dataset was created, a total of 133 elemental features encoding elemental and electronic properties were generated. This feature set is known as Magpie and was first introduced in [37] and further refined in a subsequent paper [38]. This feature set has been used in a number of works applying ML methods to various materials systems. These were in addition to the lattice parameter – the sole structural descriptor.

To select the most relevant features, we used three distinct feature selection methods: a variance threshold to remove any zero-variance features, and a combination of mutual information score with random forest permutation importance to select the 12 most important features. To find the optimal features, an exhaustive search for the best performing model within this feature space was performed.

The selected features were then used to create different Random Forest models for predicting T_c values of the compounds in the dataset. It has to be noted that to estimate the influence of the choice of a model on the results, we used a similar workflow but utilizing two other tree-based ML algorithms, namely XGBoost and Extra Trees regressors. The results were not significantly different, both regarding the selected features and the ML model accuracy (as well as the different series classified by performance groups). We have used Random Forest as it is the simplest of the three and requires the least amount of hyper-parameter tuning.

Two different approaches for splitting the data into training and testing sets were employed. As an initial ML approach, a conventional random train-test split was carried out, providing an overall good performance (R^2 of 0.82). The trained model was used to calculate T_c at several compositions and predicted some intriguing behavior which can be experimentally tested. As a more challenging task for the ML approach, we developed a novel testing methodology in which data splits based on entire doping series is used. The performance of the models on this task is more varied, with some series being predicted rather accurately, while for others the ML model shows limited predictive power. In order to examine these differences in the ML performance, the doping series were classified into Groups 1, 2 and 3 (Table 4), in the order of decreasing performance. An anomaly detection algorithm was used as a way to distinguish these groups, thus providing a way to anticipate the performance of the ML model on substitution series not represented in the training data.

The utility of this novel testing approach clearly extends beyond TM carbides and nitrides. It can be employed to benchmark ML models for materials in which chemical substitutions are used to modify their physical properties. In the field of superconductivity this clearly includes some of the most prominent classes of materials such as cuprates and iron-based compounds.

CRediT authorship contribution statement

Houssam Metni: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Ichiro Takeuchi:** Conceptualization, Funding acquisition, Writing – review & editing. **Valentin Stanev:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset and examples of the code used in this work are accessible at: https://github.com/hmetni/Superconducting_TM_C_N

Acknowledgments

The work at the University of Maryland, United States was supported by AFOSR FA9550-22-10023 and DOE STTR DE-SC0021599.

References

[1] S.T. Oyama, Introduction to the chemistry of transition metal carbides and nitrides, in: The Chemistry of Transition Metal Carbides and Nitrides, Springer, 1996, pp. 1–27.
[2] C. Kral, W. Lengauer, D. Rafaja, P. Ettmayer, Critical review on the elastic properties of transition metal carbides, nitrides and carbonitrides, J. Alloys Compd. 265 (1–2) (1998) 215–233.
[3] A. Lausche, J. Schaidle, N. Schweitzer, L. Thompson, 7.14 - Nanoscale carbide and nitride catalysts, in: J. Reedijk, K. Poeppelmeier (Eds.), Comprehensive Inorganic Chemistry II, second ed., Elsevier, Amsterdam, 2013, pp. 371–404.
[4] R.B. Levy, M. Boudart, Platinum-like behavior of tungsten carbide in surface catalysis, Science 181 (4099) (1973) 547–549.
[5] I. Parkin, A. Kafizas, 2.17 - Exothermic metathesis reactions, in: J. Reedijk, K. Poeppelmeier (Eds.), Comprehensive Inorganic Chemistry II (Second Edition), second ed., Elsevier, Amsterdam, 2013, pp. 471–490.
[6] A. Santhanam, Application of transition metal carbides and nitrides in industrial tools, in: The Chemistry of Transition Metal Carbides and Nitrides, Springer, 1996, pp. 28–52.
[7] Y. Zhong, X. Xia, F. Shi, J. Zhan, J. Tu, H.J. Fan, Transition metal carbides and nitrides in energy storage and conversion, Adv. Sci. 3 (5) (2016) 1500286.
[8] D.W. Flaherty, R.A. May, S.P. Berglund, K.J. Stevenson, C.B. Mullins, Low temperature synthesis and characterization of nanocrystalline titanium carbide with tunable porous architectures, Chem. Mater. 22 (2) (2010) 319–329.
[9] U. Guler, V.M. Shalaev, A. Boltasseva, Nanoparticle plasmonics: going practical with transition metal nitrides, Mater. Today 18 (4) (2015) 227–237.
[10] W. Ziegler, R. Young, Studies of compounds for superconductivity, Phys. Rev. 90 (1) (1953) 115.
[11] N. Pessall, J.K. Hulm, Superconducting alloys of interstitial compounds, Phys. Physique Fizika 2 (1966) 311–328.
[12] C. Yen, L. Toth, Y. Shy, D. Anderson, L. Rosner, Superconducting H c-J c and T c Measurements in the Nb–Ti–N, Nb–Hf–N, and Nb–V–N ternary systems, J. Appl. Phys. 38 (5) (1967) 2268–2271.
[13] N. Pessall, R. Gold, H. Johansen, A study of superconductivity in interstitial compounds, J. Phys. Chem. Solids 29 (1) (1968) 19–38.
[14] M. Gurvitch, J. Remeika, J. Rowell, J. Geerk, W. Lowe, Tunneling, resistive and structural study of NbN and other superconducting nitrides, IEEE Trans. Magn. 21 (2) (1985) 509–513.
[15] J. Geerk, G. Linker, R. Smithey, Electron tunneling into superconducting ZrN, Phys. Rev. Lett. 57 (26) (1986) 3284.
[16] S.D. Brorson, A. Kazeroonian, J.S. Moodera, D.W. Face, T.K. Cheng, E.P. Ippen, M.S. Dresselhaus, G. Dresselhaus, Femtosecond room-temperature measurement of the electron-phonon coupling constant γ in metallic superconductors, Phys. Rev. Lett. 64 (1990) 2172–2175.
[17] B. Wang, K. Matsubayashi, Y. Uwatoko, K. Ohgushi, High pressure effect on the superconductivity in VN, J. Phys. Soc. Japan 84 (10) (2015) 104706.
[18] B.M. Klein, D.A. Papaconstantopoulos, Electron-phonon interaction and superconductivity in transition metals and transition-metal carbides, Phys. Rev. Lett. 32 (21) (1974) 1193.
[19] E. Isaev, R. Ahuja, S. Simak, A. Lichtenstein, Y.K. Vekilov, B. Johansson, I. Abrikosov, Anomalously enhanced superconductivity and ab initio lattice dynamics in transition metal carbides and nitrides, Phys. Rev. B 72 (6) (2005) 064515.
[20] E. Maksimov, S. Ebert, M. Magnitskaya, A. Karakozov, Ab initio calculations of the physical properties of transition metal carbides and nitrides and possible routes to high-T c superconductivity, J. Exp. Theor. Phys. 105 (3) (2007) 642–651.
[21] J. Noffsinger, F. Giustino, S.G. Louie, M.L. Cohen, First-principles study of superconductivity and Fermi-surface nesting in ultrahard transition metal carbides, Phys. Rev. B 77 (18) (2008) 180507.
[22] E. Maksimov, S. Wang, M. Magnitskaya, S. Ebert, Effect of high pressure on the phonon spectra and superconductivity in ZrN and HfN, Supercond. Sci. Technol. 22 (7) (2009) 075004.
[23] N. Armitage, P. Fournier, R. Greene, Progress and perspectives on electron-doped cuprates, Rev. Modern Phys. 82 (3) (2010) 2421.
[24] B. Keimer, S.A. Kivelson, M.R. Norman, S. Uchida, J. Zaanen, From quantum matter to high-temperature superconductivity in copper oxides, Nature 518 (7538) (2015) 179–186.

[25] R.M. Fernandes, A.I. Coldea, H. Ding, I.R. Fisher, P. Hirschfeld, G. Kotliar, Iron pnictides and chalcogenides: a new paradigm for superconductivity, Nature 601 (7891) (2022) 35–44.
[26] V. Stanev, K. Choudhary, A.G. Kusne, J. Paglione, I. Takeuchi, Artificial intelligence for search and discovery of quantum materials, Commun. Mater. 2 (1) (2021) 1–11.
[27] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, Machine learning modeling of superconducting critical temperature, Npj Comput. Mater. 4 (1) (2018) 29.
[28] K. Hamidieh, A data-driven statistical model for predicting the critical temperature of a superconductor, Comput. Mater. Sci. 154 (2018) 346–354.
[29] S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, J. Ni, Atom table convolutional neural networks for an accurate prediction of compounds properties, NPJ Comput. Mater. 5 (1) (2019) 1–7.
[30] K. Matsumoto, T. Horide, An acceleration search method of higher Tc superconductors by a machine learning algorithm, Appl. Phys. Express 12 (7) (2019) 073003.
[31] Z.-L. Liu, P. Kang, Y. Zhu, L. Liu, H. Guo, Material informatics for layered high-TC superconductors, APL Mater. 8 (6) (2020) 061104.
[32] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, A. Maeda, Deep learning model for finding new superconductors, Phys. Rev. B 103 (1) (2021).
[33] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, et al., Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery, Mol. Syst. Des. Eng. 3 (5) (2018) 819–825.
[34] National Institute for Materials Science, Superconducting material database(supercon).
[35] A. Rohatgi, Webplotdigitizer: Version 4.5, 2021.
[36] W.T. Ziegler, R.A. Young, Studies of compounds for superconductivity, Phys. Rev. 90 (1953) 115–119.
[37] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, NPJ Comput. Mater. 2 (1) (2016).
[38] L. Ward, A. Dunn, A. Faghaninia, N. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. Persson, G. Snyder, I. Foster, A. Jain, Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.
[39] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (6) (2004) 066138.
[40] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (6) (1933) 417–441.
[41] L. van der Maaten, G. Hinton, Viualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[42] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.
[43] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[44] 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, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[45] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, 2008, pp. 413–422.