Computational Materials Science 171 (2020) 109203

Contents lists available at ScienceDirect

Computational Materials Science

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

Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation

Zheng Xiong[a], Yuxin Cui[a], Zhonghao Liu[a], Yong Zhao[a], Ming Hu[b], Jianjun Hu[a,c,*]

[a] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
[b] Department of Mechanical Engineering, University of South Carolina, Columbia, SC 29208, USA
[c] Department of Mechanical Engineering, Guizhou University, Guiyang 550025, China

ARTICLE INFO

Dataset link: https://github.com/buptxz/kmFCV

Keywords:
Materials discovery
Machine learning
Cross-validation
Extrapolation
Exploration
Interpolation
Performance evaluation

ABSTRACT

The materials discovery problem usually aims to identify novel “outlier” materials with extremely low or high property values outside of the scope of all known materials. It can be mapped as an explorative prediction problem. However, currently the performance of machine learning algorithms for materials property prediction is usually evaluated via k-fold cross-validation (CV) or holdout-test, which tend to over-estimate their explorative prediction performance in discovering novel materials. We propose k-fold-m-step forward cross-validation (kmFCV) as a new way for evaluating exploration performance in materials property prediction and conducted a comprehensive benchmark evaluation on the exploration performance of a variety of prediction models on materials property (including formation energy, band gap, and superconducting critical temperature) prediction with different materials representation and machine learning algorithms. Our results show that even though current machine learning models can achieve good results when evaluated with traditional CV, their explorative power is actually very low as shown by our proposed kmFCV evaluation method and the proposed exploration accuracy. More advanced explorative machine learning algorithms are strongly needed for new materials discovery.

1. Introduction

A common research problem in materials science is to discover new materials with higher or lower physical/chemical properties based on all known materials. This includes the efforts seeking materials with higher thermal conductivity [1], ionic conductivity for the fuel cell or lithium-battery materials [2,3], higher electronic conductivity, optical property, and higher superconducting critical temperature [4]. On the other hand, researchers have also been seeking insulation materials with extremely lower thermal conductivity [5,6] or materials with lower electronic conductivities, etc. The common challenge among these problems is how to identify new materials whose figure of merit or performance measure is beyond domain of all known materials. From the perspective of data science, here we are more interested in building predictive models that have stronger explorative power rather than interpolation power so that it allows us to find materials with “outlier” performance beyond region of known materials where no known samples exist. Accordingly, machine learning (ML) algorithms with high explorative power are needed to build prediction models for high-throughput screening.

In the past several years, many researchers applied data-driven based machine learning techniques to predict material properties [4,7–18], using the large-scale data collection based on computation by Density Functional Theory (DFT)[19], such as the Materials Project (MP)[20], Open Quantum Materials Database (OQMD)[21,22], the Automatic Flow of Materials Discovery Library (AFLOWLIB.ORG)[23]. However, most of current practice of machine learning in materials informatics has inappropriately stuck to the traditional machine learning model evaluation approaches [18,24,25,26]. Meredig et al. [24] summarized the reported regression performance of six materials property prediction models in the literature, all showing excellent R^2 scores while *“materials discovery has not been revolutionized yet”* and these models are far from being able to be used as a one-shot high-throughput screening of large numbers of materials for desired properties. In Table 1, a more comprehensive summary of reported model performance is shown, which are uniformly excellent across different studies on prediction models of different materials properties.

Meredig et al. [27] suggested that the traditional cross-validation (CV) has critical shortcomings in terms of quantifying ML model performance for materials discovery. Actually, many of the good

[*] Corresponding author at: Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA.
E-mail address: jianjunh@cse.sc.edu (J. Hu).

https://doi.org/10.1016/j.commatsci.2019.109203
Received 16 July 2019; Received in revised form 11 August 2019; Accepted 12 August 2019
Available online 27 August 2019

Table 1
Materials informatics model results from the literature. Mean absolute error (MAE), root mean squared error (RMSE) and coefficient of determination (R^2) are three common regression performance metrics.

Material dataset	Dataset size	Property	Technique	Evaluation method	Performance	Ref
Inorganic compounds (Materials Project)	28,046 16,458	Formation energy Band gap	Crystal Graph Convolutional Neural Network (CGCNN)	60% train, 20% validation, 20% test	MAE = 0.039 eV/atom MAE = 0.388 eV	[12]
Inorganic compounds (OQMD)	256,673	Formation energy	MLP with one-hot composition representation	90% train, 10% test	MAE = 0.072 eV/atom	[15]
Inorganic compounds (ICSD subset in OQMD)	Over 30 k	Formation energy	Random forest with Magpie descriptor	30,000 for train, 1,000 for validation	R = 0.988 MAE = 0.09 eV/atom RMSE = 0.15 eV/atom	[11]
X_2YZ formula from OQMD	69,710	Formation energy	CNN with periodic Table Representation	65,710 for train, 4,000 for test	MAE = 0.007 eV/atom	[13]
ABC_2D_6formula	About 10 k	Formation energy	MLP with Atom2Vec descriptor	Not reported	MAE = 0.15 eV/atom	[38]
SuperCon	About 9 k	Superconducting critical temperature	Random forest with Magpie descriptor	85% train, 15% test	R^2 = 0.88	[4]

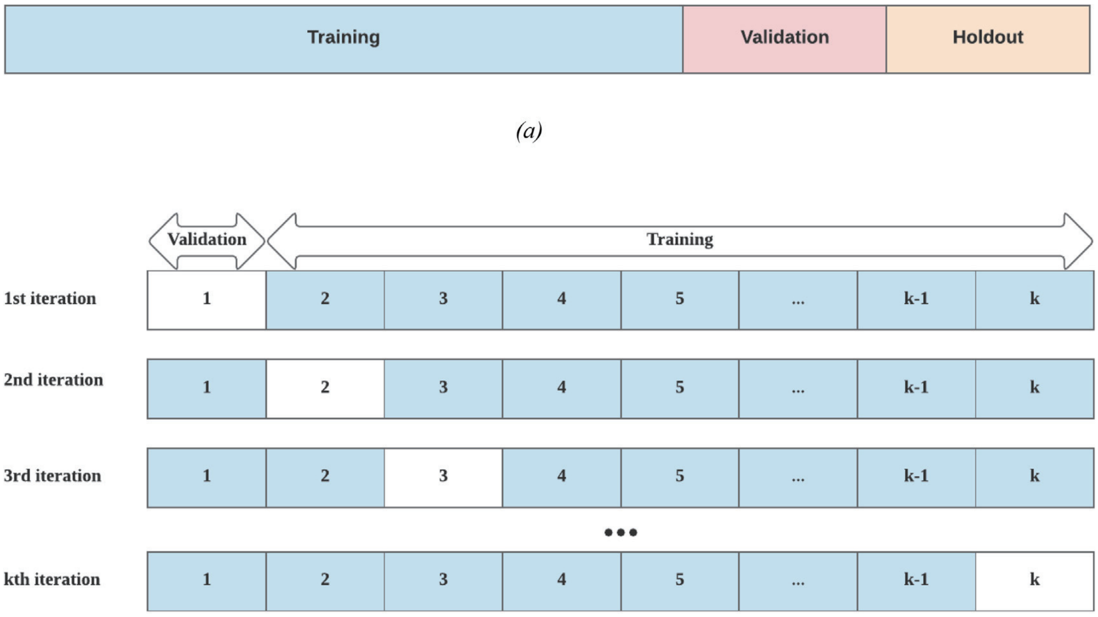
performance scores are due to the highly redundant or similar training samples in the dataset [18]. These models are very likely to fail when used to find “outlier” materials with few known samples around. New performance measures are needed to more objectively evaluate the exploration rather than interpolation power for the discovery of “outlier” materials. Indeed, few methods have been proposed to explicitly train explorative machine learning models as done in [27–29].

In standard machine learning, three types of evaluation methods are commonly used (see Fig. 1). The first one is the holdout method which randomly divides the whole dataset into a training set, a validation set and a holdout/test set, then trains the predictive model over the training set, finds the best parameters for the model over the validation set and evaluates its performance on the holdout/test set. However, the performance can be biased based on the splitting if the dataset is small. The second commonly used method is k-fold cross-validation, in which the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the validation/test set and the other $k-1$ subsets are put together to form a training set. The error estimation is averaged over all k trials to get the total effectiveness of the model. Here each sample gets to be in a validation/test set exactly once and gets to be in a training set $k-1$ times. It significantly reduces bias as we are using most of the data for fitting, and significantly reduces variance as most of the data is also being used in the validation/test set. The last method for performance evaluation is leave-one-out cross-validation, which is just a special case of k-fold cross-validation when k is set as the number of samples of the whole dataset. In this case, each time a single sample is held out as the validation/test sample while all the remaining samples are used in training.

The problem with the standard model evaluation methods is that they are designed for evaluating interpolation power rather than explorative power as needed for materials property prediction model evaluation.

2. Related work

The first investigation of the limitation of the traditional cross-validation approach for quantifying materials prediction models is reported by Meredig et al. [27]. They suggested that traditional cross-validation has critical shortcomings in terms of quantifying ML model performance for materials discovery. Basically, materials informatics practitioners prefer the extrapolative power of the model to find “outlier” samples with unseen extremely good performance. In addition,



(a)

(b)

Fig. 1. Commonly used evaluation methods in machine learning. a) hold-out method; b) k-fold cross-validation; leave-one-out as a special case of k-fold cross-validation when k is the number of samples.

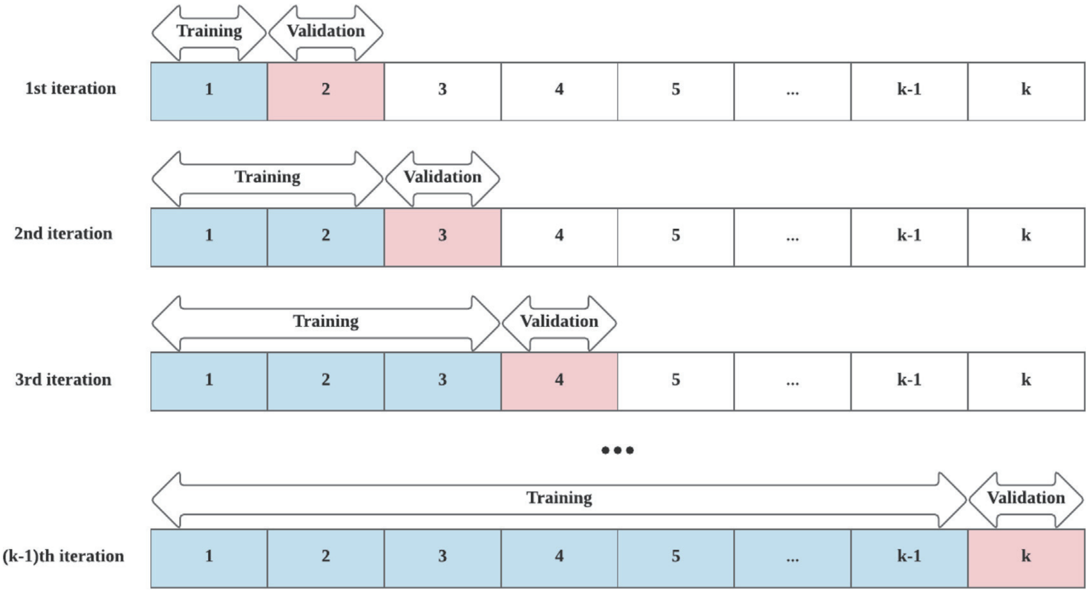


Fig. 2. Diagram of k-fold forward cross-validation.

researchers in materials science often perform extrapolative regression rather than pattern matching classification based on datasets that are neither uniformly nor randomly sampled within their domains. They proposed Leave-one-cluster-out cross-validation (LOCO CV) method to avoid the inflated performance of traditional cross-validation approach due to redundant samples from families of materials and the random sampling procedure. In their approach, instead of using random splitting of the dataset for cross-validation, LOCO CV first applies a clustering procedure to split the dataset into k clusters, and then does traditional k-fold cross-validation. This cycle will be repeated for several k values then the results across different values of k are summarized by computing median and standard deviation across k. However, a major drawback of this approach is that in real-world materials discovery, the datasets are not given by clusters. Instead, the materials discovery problem is: given all known materials of the best Figure of Merit (FOM) value of θ, can the prediction model find materials with higher FOM? This inspires us to use FOM for splitting instead of using clusters. Extrapolation power of predictive models is usually understood in terms of input domain (how the materials are represented): we have extrapolation when one asks a trained model to predict values for input data points (materials) that are outside the observation domain. Here, we are more interested in evaluating how the models can discovery new materials with “out-of-the-boundary” FOM properties. We call this prediction power as explorative prediction power compared to extrapolative prediction power.

Other related work on extrapolation of machine learning models is reported in [28,29]. Margius and Lampert [28] proposed a neural network approach to learn interpretable functions using an end-to-end differentiable feedforward neural network framework with efficient gradient-based training. The model has the advantage of being able to extrapolate to an unseen domain. The concise interpretable mathematical expression is obtained via a sparsity regularization of the loss function. Compared to interpolation of black-box regression, their approach allows understanding functional relations and generalizing them from observed data to unseen parts of the parameter space. Analytical function learning is also called symbolic regression in evolutionary computation. Schimit and Lipson [30] used genetic programming, a special form of genetic algorithms to evolve and discover a series of scientific laws from observation data. However, they have not systematically explored its performance in terms of exploration in the unseen domain.

Extrapolating in the data domain implies that the data distribution at prediction time will differ from the data distribution at training time. LOCO CV addresses the limitation of the traditional cross-validation method by splitting the samples in the feature space, which gives a better evaluation measure of the explorative power of the predictive models. This paper aims to develop a set of new evaluation methods – k-fold-m-step forward cross-validation (kmFCV) – aiming to split the training samples according to the property values and to evaluate how likely a predictive model can predict the materials property outside of the training samples domain. This method avoids the arbitrary determination of the clusters needed in LOCO CV approach. Our main contributions can be summarized as:

1. We propose a set of new forward cross-validation methods and a new metric for evaluating the explorative prediction power of materials property prediction models.
2. We compare how our forward cross-validation methods help to differentiate high explorative models from low ones.
3. We evaluate how materials representation/descriptor and machine learning algorithm affect the explorative prediction power of machine learning models through extensive benchmark studies.

The remaining part of the paper is organized as follows: Section 3 introduces the explorative forward cross-validation methods, the benchmark datasets, benchmark materials descriptors, and benchmark machine learning algorithms. Section 4 shows the comparison results of the proposed forward cross-validation methods and traditional cross-validation method. We discuss the advantages and disadvantages of our method and how different factors affect the explorative power of the materials property prediction models. We conclude our paper in Section 5.

3. Methods

3.1. New evaluation methods for benchmarking the explorative prediction capability

In order to evaluate if a prediction model of material properties has exploration capability, i.e. trained on a group of materials and has predictive power for materials in a different domain. We propose the following set of explorative evaluation schemes:

3.1.1. Forward holdout validation (Forward-holdout)

This evaluation method is similar to the holdout validation method in standard machine learning except that we first sort all data samples by the ascending/descending target property and then split it into training and validation sets. When sorted by ascending, the subset with lower property values is set for training, and the other set is used for testing. This evaluation method is good for learning models to discover materials with higher property values. Similarly, if the dataset is sorted by descending, the set with higher property values is set as the training set, and the other set is used for testing. This evaluation method is good for learning models to discover materials with lower property values such as extremely low thermal conductivity.

3.1.2. K-fold forward cross-validation (k-fold FCV)

The k-fold forward cross-validation is an improvement over traditional k-fold cross-validation for evaluating explorative prediction power of models. Instead of randomly partitioning the dataset, all the samples are first sorted by the materials property values and then split into k subsets evenly. The whole process is as follows:

1. Sort all samples by ascending/descending property values
2. The sorted samples are partitioned into k equal-sized subsets S_1, S_2, …, S_k
3. Starting from the second subset S_2, set S_2 as the validation set and the first subset S_1 as the training set, train a model on S_1, and evaluate its performance on S_2
4. Next round, set S_3 as the validation set and all subsets before S_3 as the training data, i.e., train a model on S_1 and S_2 and evaluate its performance on S_3
5. Repeat step 4 until all S_2 to S_k have been evaluated. Calculate the overall performance of all models

Whether to sort the samples by ascending or descending property values depends on which side – higher or lower – we expect the model to extrapolate. With the relatively large k value, the training set size might be too small at the very first beginning. A minimum size of the training data can be set to prevent the result from being distorted by this issue.

3.1.3. Leave-one-out forward cross-validation (LOOFCV)

Leave-one-out forward cross-validation (LOOFCV) is a special case of k-fold forward cross-validation in which the samples are split into N subsets where N is the total number of samples.

3.1.4. K-fold-m-step forward cross-validation (k-fold-m-step FCV)

K-fold-m-step forward-cross-validation is an extension of k-fold forward-cross-validation by assuming an $m-1$ number of subset gap between the last training subset and the validation subset. Starting with the $(m+1)$th subset set as the validation set and the first subset as the training set, train the model and evaluate its performance. Then set the $(m+2)$th subset as the validation set and the first and second subset as training sets. Repeat this until the last subset is set as the validation set. This evaluation method can be used to test how a given model can predict the materials property values m step of subset away from the training set. It can avoid the overestimated performance due to local nearby redundancy samples. k-fold forward cross-validation is a special case of k-fold-m-step forward cross-validation when m is 1.

3.1.5. Performance metrics

Similar to the case of traditional cross-validation, the performance metrics of forward cross-validation methods can be calculated from the subsets of corresponding predictions. Three different performance metrics – Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R^2) can be calculated across all experiments in order to compare the performance of different algorithms. All the prediction results of subsets will then be aggregated to calculate the overall performance.

In addition to the three traditional regression metrics, we introduce a new classification metric called *exploration accuracy* ($E_{Accuracy}$) to evaluate the exploration performance of prediction models. When doing k-fold-m-step forward cross-validation, for each fold, a training domain threshold θ is set as the largest property value in the training set of that fold. And then a sample in the validation set is labeled as positive if its predicted property value is equal to or larger than the domain threshold or as negative if the predicted value is smaller than the threshold value θ. An accuracy score can then be calculated based on all these positives and negatives: $E_{accuracy} = \frac{\#\,Positive}{\#\,Positive + \#\,Negative}$. This *exploration accuracy* metric can be used to evaluate how likely a model can correctly classify a validation sample into within-domain or outside-domain. An exploration compatible predictor should achieve high *exploration accuracy*. On the other hand, predictors that cannot predict property values outside of the domain of its training samples will get zero $E_{accuracy}$.

3.2. Benchmark datasets

We prepared three benchmark datasets from public databases for evaluating explorative prediction capabilities of current machine learning algorithms and descriptors. Firstly, the publicly available Materials Project database [20], which contains more than 86,000 inorganic compounds when the data was collected, is used in this study as the benchmark dataset to predict two materials properties – formation energy and band gap. It contains the Density Functional Theory [19] calculated properties of all these compounds. The materials in MP consist of as many as 7 different elements and over 90% of them are binary, ternary or quaternary compounds. The number of atoms varies from 1 to 200. Formation energy and band gap are selected as benchmark properties because they are the most researched properties by materials informatics researchers and the evaluation results can be compared with all previous studies.

The second benchmark dataset in this study is the SuperCon database [31] with the critical temperatures of all known superconductors. It was created and maintained by the Japanese National Institute for Materials Science. The superconducting critical temperatures are from experiments. Stanev [4] has been putting the effort in developing a classification model to classify the samples into two classes with above or below 10 K critical temperature and a regression model to predict the critical temperature using random forest (RF)[32] algorithm with Magpie descriptor [10]. We extracted a list of about 9 k samples with the superconducting critical temperature above 10 K as the exploration preformation benchmark dataset as we are looking for materials with higher critical temperatures.

Data filtering is conducted before the datasets are used for the benchmark. The MP database contains DFT calculations of formation energy and band gap for 83,989 compounds when the data were collected. In the case that more than one compounds share the same composition, the one with the lowest formation energy is chosen since lower formation energy indicates a more chemically stable compound. Also, outliers of which the property values are outside of $\pm 5\sigma$ (σ is the standard deviation) bound are removed. The compounds with only one element are also removed as their formation energy is considered as zero. What's more, ill converged samples in the MP dataset – (a) any crystal with a warning tag, which usually states a significant structural change during relaxation, (b) any crystal without a calculated band diagram, since it usually means the calculation is not very accurate – are removed. After the filtering, there remain 35,216 compounds out of MP for formation energy prediction evaluation, 20,065 compounds out of MP for band gap prediction evaluation, and 6,258 compounds out of SuperCon for superconducting critical temperature prediction evaluation as shown in Table 2.

When comparing different representation/descriptor performance,

Table 2
Datasets summary. The dataset sizes and names before and after data filtering.

Dataset	Original size	Complete Set	Representation Set
Formation energy (MP)	83,989	35,216 (MPFE-35 K)	18,274 (MPFE-18 K)
Band gap (MP)	83,989	20,065 (MPBG-20 K)	10,042 (MPBG-10 K)
SuperCon	16,413	6,258 (SC-6 K)	2,876 (SC-2.8 K)

the dataset needs to be filtered to be applicable to all of them in order to prevent introducing additional factors. The Periodic Table Representation (PTR)[13], which will be briefly introduced later, can only represent the compound with 52 elements in the periodic table due to its limitation. The compounds that contain elements out of these 52 ones are removed. After removing the compounds out of the PTR representation elements, the representation set sizes are 18,274, 10,042 and 2,876 respectively, shown in Table 2.

3.3. Materials descriptors and machine learning algorithms for materials property prediction

We evaluated the exploration performance of a set of machine learning algorithms with the following materials descriptors: Magpie, elemental one-hot composition representation, period table representation (PTR), and crystal graph representation.

Magpie (Materials Agnostic Platform for Informatics and Exploration) is a materials descriptor proposed by Ward et al. to convert compound composition into meaningful features [10], which can be conveniently calculated using matminer [33] and Pymatgen [34]. It computes a set of features for a given material including elemental property statistics like the mean and the standard deviation of 22 different elemental properties.

One-hot composition representation [15] is an encoding method for compounds based on their formula. The size of the feature vector is set to the total number of elements in the dataset and the feature values are the ratios of the constituent elements in the compound. For example, for a compound with the formula CO_2, the values of the corresponding positions of C and O in the feature vector are 0.33 and 0.66 respectively, with all other positions being set as zero and all nonzero values being added up to be 1.

Periodic table representation [13] uses a 2-D matrix to represent the composition of a compound, which makes it similar to an image consisting of 2-D pixels. Pattern recognition of images has been widely and successfully addressed using convolutional neural networks (CNNs). In this encoding, all values of the matrix are first initialized as -1 and then the blank spaces in the corresponding spot of the periodic table are set as 0. The values of the element positions in the matrix are set as the atom numbers of the elements in the compound. Then the matrix is multiplied by 20 to mimic a digital image to ease the training process of the CNN. Due to the limitation of the representation, only a rectangular area of the periodic table can be represented, which means the compounds with the elements outside of this area cannot be represented. CNN with PTR has been applied to an X_2YZ formula materials group for formation energy and stability prediction with great performance by Zheng et al. [13], as shown in Table 1.

Crystal graph representation of compounds is used in CGCNN [12], an end-to-end graph convolution based deep learning framework for materials property prediction. It is the only model we evaluated here that utilizes the crystal structure information for property prediction, which directly learns material properties from the interactions of atoms in the crystal, providing a universal and interpretable representation of crystal structures. Since the SuperCon dataset doesn't come with structural information, CGCNN can only be applied to the MP datasets. This model has been applied to predict 7 properties over the MP dataset and demonstrated the superior performance compared to the other models tested here, as shown in Table 1. The convolutional neural model consists of three convolutional layers with 64 nodes followed by a fully connected layer with 128 nodes. SGD is used for optimization. Batch size is 256 and 30 epochs are trained for each model.

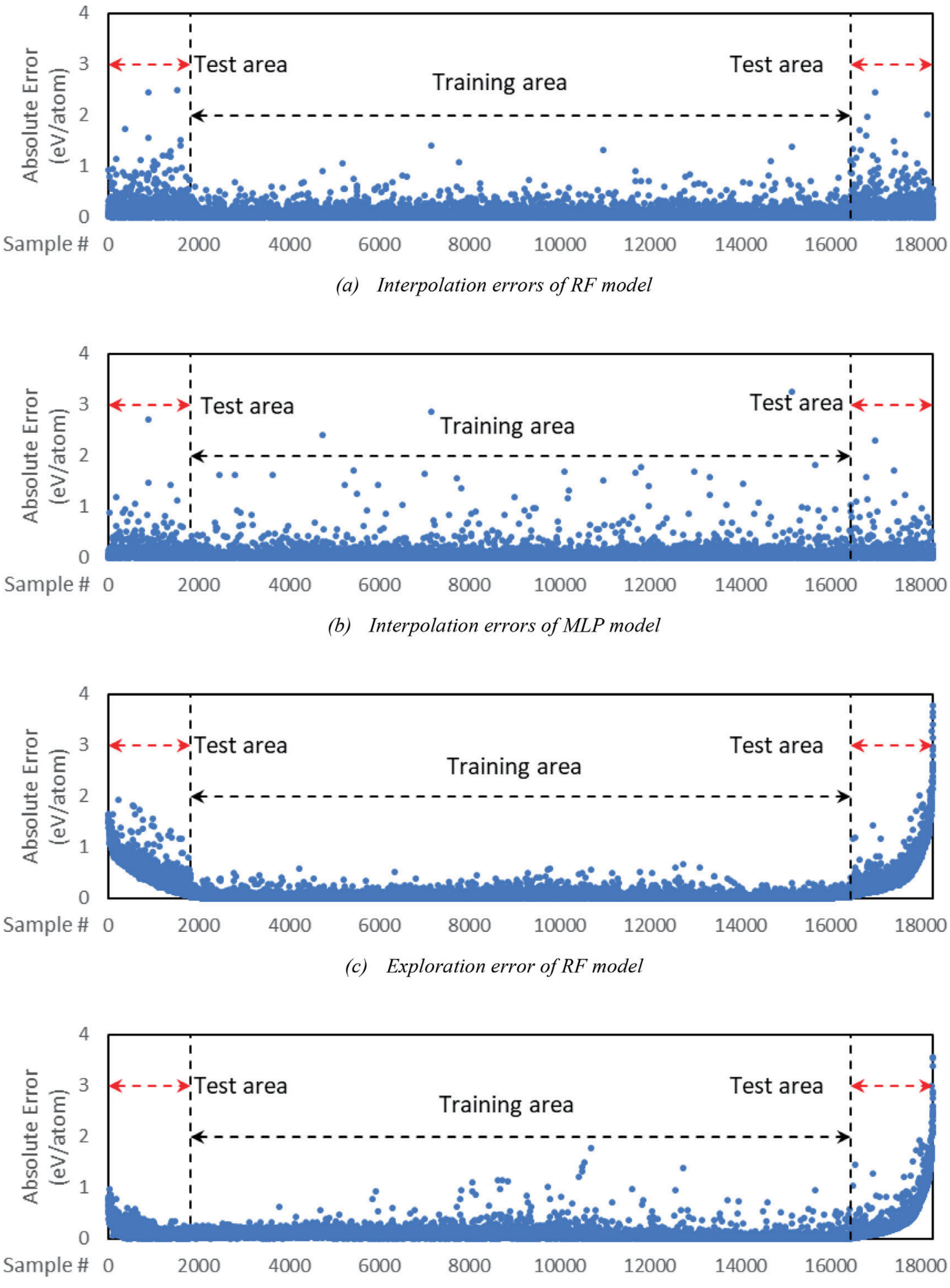
Three types of machine learning models are used in our experiments to build materials property predictors using the Magpie descriptor and one-hot composition representation, both of which are one-dimensional feature vectors. They are 1-Nearest-Neighbor (1NN), random forest (RF) and multilayer perceptron neural network (MLP). First, a naive 1NN has been suggested as an essential benchmark to contextualize performance of materials informatics models [27]. It serves as the baseline for the property prediction problem, as the prediction is based on the nearest sample in terms of a distance metric (Euclidean distance used in our model) from the training set. Since test samples in exploration experiments are usually regarded as outliers, little explorative power is expected from 1NN models. Second, RF is a popular ensemble model that has been used in a variety of materials property prediction research [4,10,11]. It is a bagging technique that builds up a strong learner from an ensemble of weak decision trees. Both 1NN and RF are implemented using the scikit-learn python package [35]. The number of decision trees for RF is 100, which comes from the experience of the previous research [11]. The max number of features in RF model is 10, in order to significantly reduce the calculation time but maintain a relatively good performance, as the k value is large (see Section 4.2 for justification of approach). The third model is MLP, which has been well investigated and the strong prediction power for formation energy was reported using one-hot composition representation [15]. MLP is a vanilla artificial neural network which usually consists of three or more fully connected layers with nonlinear activation functions. A seven-layer MLP with six fully connected layers with 1024, 1024, 512, 512, 128, 128 nodes and an output layer with 1 node, is trained based on the hyper-parameter tuning experiments from previous research. We used a batch size of 128 with Adam [36] optimizer in training. There are three dropout layers with 0.5 rate between fully connected layers with different nodes. The loss function is MAE and a total number of training epochs is set as 30.

4. Results and discussion

4.1. Why a new evaluation method is needed to measure exploration performance for materials property prediction

Machine learning models are usually evaluated by cross-validation to measure the interpolation power. However, in order to drive materials discovery for superior properties, good exploration performance is the key to enabling the ML models to be used for high-throughput screening of materials with desired properties. However, traditional cross-validation methods in which the samples are randomly split into folds are not well suited for evaluating the exploration performance.

To illustrate this issue, we did a special training/test split on the MP formation energy dataset and applied two models to it – RF and MLP with the same one-hot composition representation. The samples are first randomized and then split into three sets – first 10%, middle 80%, and last 10%. The middle set is used as the training set and the two end sets are used as the test sets. Fig. 3(a) and (b) shows the prediction errors of RF and MLP using this special hold-out method. The absolute errors of all the samples are shown as scattered dots. The difference between RF and MLP can hardly be observed.



(a) Interpolation errors of RF model

(b) Interpolation errors of MLP model

(c) Exploration error of RF model

(d) Exploration error of MLP model

Fig. 3. Illustration of interpolation and exploration errors for RF and MLP models with one-hot composition representation tested on the MP formation energy dataset. The samples are split into 10%, 80% and 10% distribution according to the sample order. The middle set is used as the training set and then the models are evaluated over two test sets from both ends. However, in (a) and (b) the samples are unsorted and randomized, while in (c) and (d) the samples are sorted by the formation energy value.

In Fig. 3(c) and (d), however, the training and test set split method is the same, but they are sorted by the ascending order of the formation energy values instead of being randomized. In this way can we evaluate how the trained RF and MLP models extrapolate from training samples to samples with higher or lower properties. Both models achieve very low errors in the middle training area like the previous experiment but don’t perform well in the test area. As the property value of the test samples deviates more from the training set boundary, the prediction errors increase dramatically. However, compared with the previous experiment, the difference between RF and MLP becomes much more visible. The MLP model (see Fig. 3(d)) generalizes better than RF (see Fig. 3(c)), especially in the left end test area. That’s why we have been working on developing a new evaluation method to measure exploration performance and analyzing how we are able to improve the explorative power of current ML models.

4.2. *K-fold forward cross-validation helps to differentiate machine learning models in terms of their explorative power*

To compare traditional cross-validation and our forward cross-validation, we need to set the parameter k, the number of folds. In forward cross-validation, the k value has a big impact on the result as it decides how different in property values the validation set will be from the training set. For traditional cross-validation, the k value would usually be set as 5 to 10. When using forward cross-validation as the evaluation method, however, we cannot expect the materials property prediction models to perform well when predicting the unseen samples that are far away from the domain of the training set. Therefore, larger k makes more sense. The k value should be set the same for traditional cross-validation and forward cross-validation, in order to compare the results of both methods in the same circumstance. The problem that remains is to find out how the k value influences the forward cross-validation results. A series of CV and FCV experiments using $k = 10$, 50, 100, 200, 500 are conducted to predict the MP formation energy using RF models with Magpie descriptor. The MAE results are shown in Table 3 and Fig. 4.

As expected, the CV results don't change too much with different k, because the k value doesn't have a large impact on the distribution of training and validation set due to the random split of samples. For FCV, however, the MAE decreases significantly with increasing k when k is small but the rope becomes smoother when k is larger than 100, as shown in Fig. 4. The tradeoff is between the accurate explorative power evaluation and computational intensity. We set k to 100 for all the following experiments as the FCV results are stable when k is larger than 100.

Table 4 shows all the property prediction results, using four metrics: MAE, RMSE, R^2, and $E_{accuracy}$ which is exclusively for FCV. The MAE comparisons between two evaluation methods and $E_{accuracy}$ of FCV are plotted in Fig. 5, providing a more intuitive view of the results.

In order to verify the correctness of our benchmark implementation of the machine learning models, we compared our evaluated CV performance with those reported in previous researches (see Table 1 and Table 4). Ward et al. [11] applied the RF with Magpie descriptor for formation energy prediction on the ICSD subset of OQMD dataset and got an MAE of 0.09 eV/atom and a RMSE of 0.15 eV/atom, which is comparable with the MAE of 0.0929 eV/atom and the RMSE of 0.1722 eV/atom in our experiments. Liu et al. [15] achieved an MAE of 0.072 eV/atom when applying MLP with one-hot composition representation to the OQMD dataset, which is also comparable with the MAE of 0.0785 eV/atom in our experiment. For SuperCon dataset, recent research [4] tried RF with Magpie descriptor and got a R^2 of 0.88, comparable with our 0.9158 result. Zheng et al. [13] trained a CNN model using PTR representation on the X_2YZ type chemical formula group from the OQMD dataset and achieved an MAE of 0.007 eV/atom. From our experiment, an MAE of 0.1085 eV/atom is obtained when applying this method to the general inorganic MP dataset. Xie and Grossman [12] proposed CGCNN for formation energy and band gap prediction and achieved MAEs of 0.039 eV/atom and 0.388 eV respectively, compared with 0.1235 eV/atom and 0.5372 eV in our experiments. The results of the last two models – CNN with PTR and CGCNN – are not quite the same when comparing with the results in the original studies (of which the CGCNN is the same implementation by Xie and Grossman). The difference might come from different dataset distributions. However, since our focus is on comparing the traditional CV and forward CV in evaluating explorative power, the identical machine learning model implementation and dataset guaranteed that the results are correct and comparable.

Table 3
Comparison of CV and FCV validation errors with different k value on formation energy prediction using RF with Magpie descriptor.

Evaluation method/k	10	50	100	200	500
CV	0.0952	0.0927	0.0925	0.0923	0.0924
FCV	0.3495	0.2146	0.1923	0.1822	0.1787

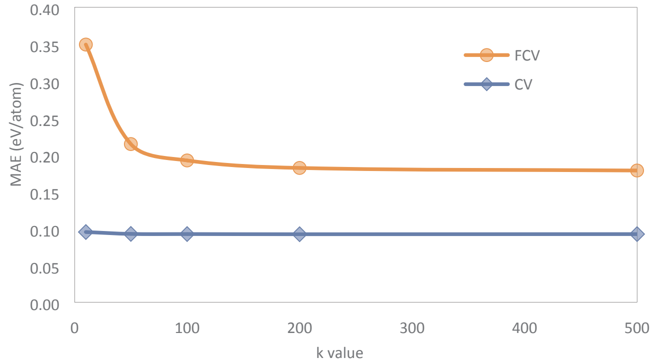


Fig. 4. Trend of MAE using different k of k-fold CV and k-fold FCV, trained on the formation energy dataset using RF with Magpie descriptor.

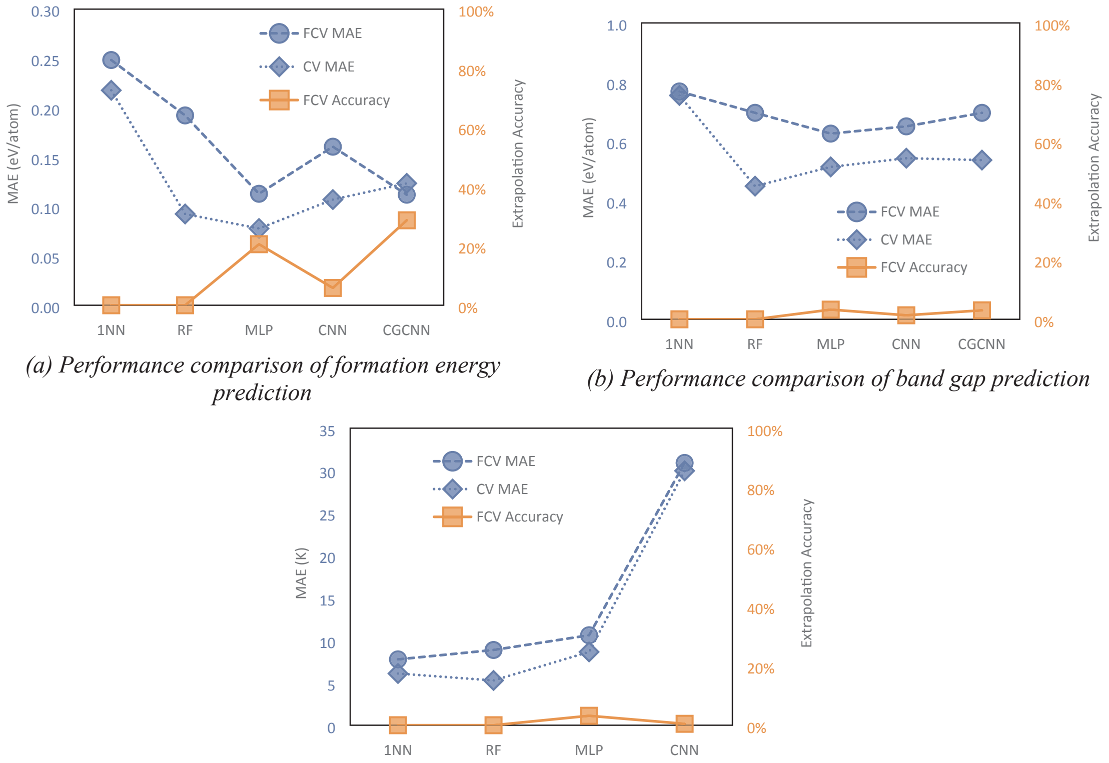
By comparing the CV and FCV results of evaluated models on predicting each of the three different materials properties in Table 4, we can have an overview of the explorative power of these benchmark models. For the remaining discussion in this part, we will use 1NN, RF, MLP, CNN and CGCNN to represent these benchmark ML algorithms with descriptor/representation models: 1NN with Magpie descriptor, RF with Magpie descriptor, MLP with one-hot composition representation, CNN with PTR and CGCNN - crystal graph CNN.

First, for the formation energy prediction problem, the best performance with traditional CV is achieved with an MAE of 0.0785 eV/atom by MLP while the worst performance belongs to 1NN with an MAE of 0.2178 eV/atom, which can be regarded as the baseline model as 1NN algorithm is one of the simplest prediction models. This shows that MLP has strong interpolation prediction capability. However, when evaluating the explorative prediction capability, the best performance is achieved by CGCNN with an FCV MAE of 0.1120 eV/atom and R^2 of 0.9658. It is impressive that in this case, the MLP still achieves a close second-best performance with FCV MAE of 0.1129 eV/atom and R^2 of 0.9582. Altogether, these results show that MLP is a strong prediction model for both interpolation and exploration for formation energy prediction. Another observation is that the FCV performance scores such as MAE, RMSE, and R^2 of the evaluated algorithms are always worse than those of the traditional CV performance, indicating the overestimated results of traditional CV evaluation methods when used to make explorative predictions. When evaluated with the exploration accuracy, it is interesting to find that both 1NN and RF achieve 0% while MLP, CNN, CGCNN achieve 20.61%, 5.84%, and 28.69% accuracy. Actually, both 1NN and RF achieved 0% explorative prediction accuracy for predicting all three properties. This shows that these two models are not able to predict property values outside the range of the training samples. While MLP, CNN, and CGCNN all have certain exploration capability, their low exploration accuracy ($< 30\%$) demonstrates that current machine learning algorithms struggle to meet the requirement for the discovery of materials with out-of-bound property values.

For the band gap prediction problem (see Table 4), the best CV performance is achieved by RF. It has an R^2 of 0.8050, which is significantly better than that of three neural network methods tested here including MLP (R^2 0.7169), CNN (R^2 0.6921), and CGCNN (R^2 0.7348). The 1NN only has a R^2 of 0.3592. However, when evaluating the explorative power, the best performance is achieved by the neural

Table 4
Summary of MAE, RMSE, R^{2}and E$_{accuracy}$ on different datasets, properties, ML algorithm with descriptor/representation models and evaluation methods. Three benchmark problems are formation energy from MP dataset, band gap from MP dataset and superconducting critical temperature from SuperCon. Both CV and FCV use 100-fold. Five ML algorithm & descriptor/representation models are Magpie descriptor with 1NN, Magpie descriptor with RF, one-hot composition representation with MLP, PTR representation with CNN, and CGCNN with crystal graph representation.

Benchmark problem	Evaluation method	Metrics	1NN with Magpie	RF with Magpie	MLP with one-hot Encoding	CNN with PTR	CGCNN -crystal graph CNN
Formation energy prediction	100-fold CV	MAE (eV/atom)	0.2178	0.0929	**0.0785**	0.1085	0.1235
		RMSE (eV/atom)	0.3641	0.1722	**0.1598**	0.2027	0.1719
		R^2	0.8833	0.9739	**0.9775**	0.9638	0.9739
	100-fold FCV	MAE (eV/atom)	0.2484	0.1923	0.1129	0.1606	**0.1120**
		RMSE (eV/atom)	0.3835	0.2468	0.1898	0.2406	**0.1555**
		R^2	0.8293	0.9293	0.9582	0.9328	**0.9658**
		$E_{accuracy}$	0%	0%	20.61%	5.84%	**28.69%**
Band gap prediction	100-fold CV	MAE (eV)	0.7553	**0.4511**	0.5156	0.5428	0.5372
		RMSE (eV)	1.1030	**0.6085**	0.7331	0.7645	0.7095
		R^2	0.3592	**0.8050**	0.7169	0.6921	0.7348
	100-fold FCV	MAE (eV)	0.7689	0.6967	**0.6266**	0.6510	0.6966
		RMSE (eV)	1.0990	0.7800	0.7663	**0.7603**	0.7985
		R^2	0.2476	0.6210	0.6342	**0.6399**	0.6028
		$E_{accuracy}$	0%	0%	**3.27%**	1.36%	3.03%
Superconductivity critical temperature prediction	100-fold CV	MAE (K)	6.0926	**5.3000**	8.6967	29.9755	N/A
		RMSE (K)	12.1576	**9.1888**	14.2092	41.6868	
		R^2	0.8526	**0.9158**	0.7987	−0.7329	
	100-fold FCV	MAE (K)	**7.7584**	8.8649	10.6022	30.9053	N/A
		RMSE (K)	14.0036	**12.8201**	14.1407	49.6371	
		R^2	0.7905	**0.8244**	0.7864	−1.6321	
		$E_{accuracy}$	0%	0%	**3.18%**	0.47%	



(a) Performance comparison of formation energy prediction

(b) Performance comparison of band gap prediction

(c) Performance comparison of superconducting critical temperature prediction

Fig. 5. Interpolation (CV) and exploration (FCV) performance comparison of five models for three materials properties: a) formation energy; b) band gap; c) superconducting critical temperature; MAEs and exploration accuracy from 100-CV and 100-FCV evaluations are reported for five different models including (1) 1NN with Magpie descriptor, (2) RF with Magpie descriptor, (3) MLP with one-hot representation, (4) CNN with PTR, (5) CGCNN-crystal graph CNN.

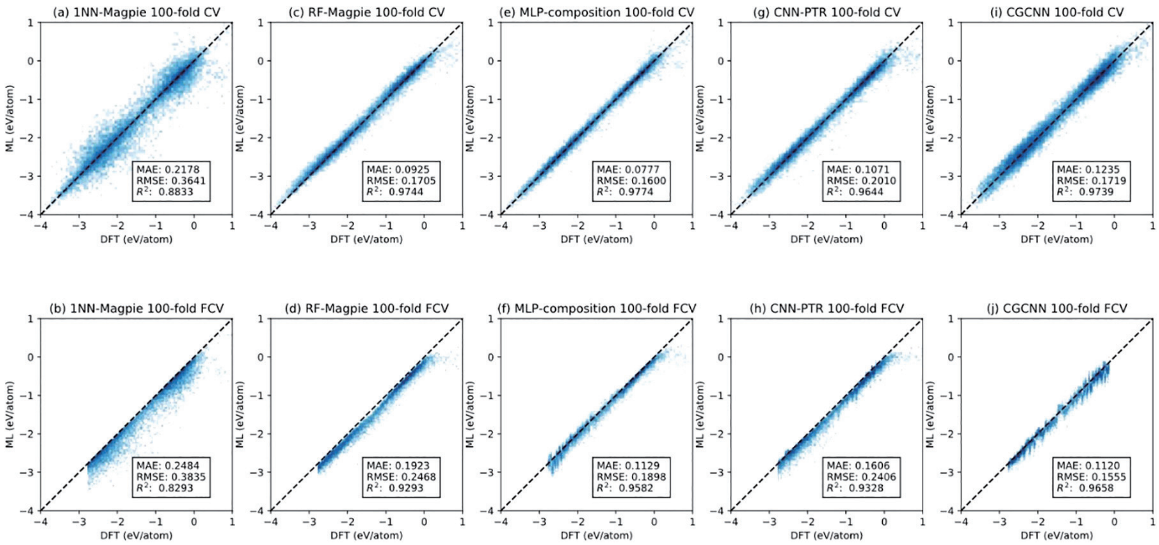


Fig. 6. Parity plots comparison of five ML models for formation energy prediction over the MP formation energy dataset. The first row and second row are the comparisons between 100-fold CV and 100-fold FCV using the same ML models. The five models are (1) 1NN-Magpie: 1NN with Magpie descriptor, (2) RF-Magpie: RF with Magpie descriptor, (3) MLP-composition: MLP with one-hot composition representation, (4) CNN-PTR: CNN with PTR, (5) CGCNN: crystal graph CNN.

network method CNN with a R^2 score of 0.6399, which is however much lower than the best R^2 score (0.9658) for formation energy prediction. In this case, both RF and 1NN get a 0% exploration accuracy score, showing their lack of explorative power. While the three neural network models have non-zero exploration accuracy, their accuracy scores prediction (3.27% for MLP, 1.36% for CNN, and 3.03% for CGCNN) are significantly lower than those for formation energy prediction (20.61% for MLP, 5.84% for CNN, and 28.69% for CGCNN). The much lower best performance metrics for band gap prediction compared to formation energy prediction shows the challenge for this problem. This unusually low exploration accuracy scores of current ML methods show that more advanced explorative prediction algorithms need to be developed for materials discovery, especially for discovering "outlier" materials with extremely low or high property values located out of the range of known materials.

For the superconducting critical temperature prediction problem, the best CV performance (see Table 4) is again achieved by RF with an R^2 of 0.9158, which is seconded by 1NN with 0.8526. By comparison, the MLP only achieves a R^2 score of 0.7987. Surprisingly, the CNN totally fails with a R^2 of -0.7329. When evaluated with FCV to measure their explorative power, the best R^2 score of 0.82 is achieved still by RF, followed by 1NN with a score of 0.7905 and MLP with a score of 0.7864. RF and 1NN still have 0% exploration accuracy while the MLP and CNN also get only 3.18% and 0.47%, both of which are much lower than the formation energy prediction problem and the band gap prediction problem. This indicates the much higher challenge for predicting superconducting critical temperature and discovering novel superconductor materials with extremely high critical temperatures. We also analyzed the main reason that 1NN performs well on this SuperCon dataset. It is partially due to this dataset containing many closely related materials varying only by small changes in stoichiometry [4], which enables 1NN to make a good prediction by identifying the neighbor sample. A similar explanation applies to RF, which is related to nearest-neighbor methods.

To provide a more comprehensive picture of the comparison between three prediction problems, we plot their CV MAE, FCV MAE and $E_{accuracy}$ in Fig. 5. First, it can be found that the FCV MAEs are all higher than the traditional CV results for all three problems, which indicates that the traditional CV tends to over-estimate the prediction performance when used to evaluate the explorative prediction powers of machine learning algorithms. The second observation is that the largest discrepancy between the traditional CV MAEs and the FCV MAEs are all from RF for all three prediction problems. For instance, RF has a CV MAE of 0.0929 eV/atom and FCV MAE of 0.1923 eV/atom when predicting formation energy, and CV MAE of 0.4511 eV and FCV MAE of 0.6967 eV when predicting band gaps. This demonstrates that even though RF works well in interpolative prediction when predicting the properties of validation samples within the same domain as the training set, its performance of exploration is not as good as neural network algorithms and is just comparable with the 1NN method.

In comparison, the other three neural network models MLP, CNN and CGCNN achieve more similar CV and FCV MAE results as shown in Fig. 5 except the case of superconductivity prediction, for which CNN does significantly worse than the other NN methods. However, the exploration accuracy of all five algorithms is still quite low, showing the difficulty of the current ML models to correctly predict the property values for samples outside of the training samples domain. There is still large room for ML models to make progress in explorative prediction.

With close inspection, we found that the lack of exploration capability of random forest algorithms is due to its underlying mechanism. Random forest is an ensemble method with multiple decision trees. The regression predictions of a random forest are done through averaging the results obtained by its multiple decision trees. However, decision trees cannot predict values outside the range of the training samples. A regression tree consists of a hierarchy of nodes, where each node specifies a test to be carried out on an attribute value and each leaf (terminal node) specifies a rule to calculate a predicted output. The trees themselves output the mean value of the samples in the leaves. It's impossible for the result to be outside the range of the training samples because the average is always inside the range of its constituents. Similarly, the 1NN algorithm can only predict the property value of a test sample as the value of the closest neighbor, which cannot be outside the range of the training samples either. Lin and Jeon [37] explains the relationship between random forests and the k-nearest neighbor algorithm (KNN) and showed that both can be viewed as weighted neighborhoods schemes. This demonstrates that exploration accuracy $E_{accuracy}$ is a good metric for measuring the explorative power of prediction models.

To further compare the traditional CV with FCV results, Fig. 6 presents the parity plots to show how the predicted formation energies are compared with the DFT calculated values (ground truth) for all five models using two evaluation methods. The figures in the first row are 100-fold CV results while the ones in the second row are 100-fold FCV results. From the 100-fold FCV results (see Fig. 6(b) and (d)), it is clear to observe that in the 1NN and RF results all the points are below the lines on which the predicted values are equal to the DFT values. This shows that both 1NN and RF cannot predict any formation energy higher than the true DFT values. Moreover, the 0% $E_{accuracy}$ of these two models means they cannot even predict beyond property values of the training set. Also, we found that the points in Fig. 6(h) are also mostly located below the line, indicating the CNN model also has low explorative power. It actually has an $E_{accuracy}$ of only 5.84%. On the other hand, the MLP and especially CGCNN have much more points located above the equation line, which correspond to their higher exploration accuracy of 20.61% and 28.69% respectively. This shows that our proposed exploration accuracy $E_{accuracy}$ is a good metric to measure if a model has good exploration capability, that is, predicting outputs beyond those of the training set.

4.3. Comparison of explorative prediction performance of machine learning algorithms

To compare the explorative power of different algorithms, we need to eliminate the influence of materials representation. Here three machine learning models including 1NN, RF and MLP are trained using the same one-hot composition representation for predicting three materials properties: formation energy, band gap, and superconducting critical temperature. The datasets of formation energy and band gap are extracted from Materials Project with 35,216 and 20,065 samples respectively. The SuperCon dataset has 6,258 samples. Compared with the experiments in Section 4.2, the datasets here are of much larger sizes.

The results are shown in Table 5 and Fig. 7. First, we found that MLP achieves the best CV performance with the lowest MAEs for both formation energy and band gap prediction problems. However, its CV performance on superconducting critical temperature prediction is the worst with an MAE of 9.0829 K compared to 5.7657 K of RF and 6.6673 K for 1NN, indicating the unusual characteristics of this problem. Next, we compared the 100-fold CV performance of the algorithms with those of 100-fold FCV as shown in Fig. 7. We found the CV performance is all over-estimated when used to measure the explorative power as the FCV is designed for. For example, the RF achieves a CV MAE of 0.1505 eV/atom for formation energy prediction while its FCV score is only 0.2839 eV/atom, almost double of the error (See Table 5). It is also found that the discrepancies between CV and FCV MAEs for the RF models are consistently the largest compared to other algorithms, which is consistent with our previous analysis, showing that RF performs well on CV (partially due to the redundant similar samples in the training and test sets) but doesn't really have explorative power.

Table 5
Exploration performance comparison of 1NN, RF and MLP algorithms for predicting formation energy, band gap and superconducting critical temperature using one-hot composition representation.

Benchmark problem	Evaluation method	Unit	1NN MAE	RF MAE	MLP MAE
Formation energy prediction	100-fold CV	eV/atom	0.2034	0.1505	**0.0719**
	100-fold FCV		0.2157	0.2839	**0.1022**
Band gap prediction	100-fold CV	eV	0.8692	0.5957	**0.5210**
	100-fold FCV		0.9412	0.8575	**0.6612**
Superconducting critical temperature prediction	100-fold CV	K	6.6673	**5.7657**	9.0829
	100-fold FCV		**7.8075**	9.3510	10.8964

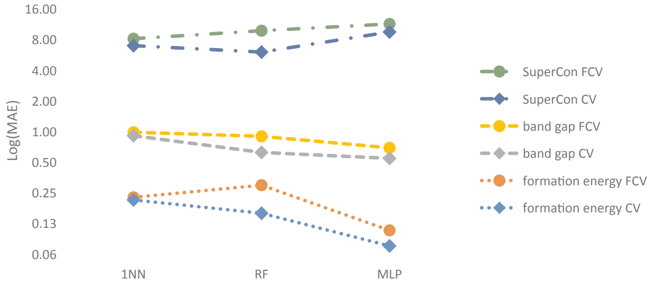


Fig. 7. CV and FCV MAE comparison of three algorithms for predicting formation energy, band gap and superconducting critical temperature using one-hot composition representation.

4.4. Explorative power evaluation using k-fold-m-step forward cross-validation

To further evaluate the explorative power of prediction models, we generalize the standard k-fold FCV to k-fold-m-step FCV. In the standard FCV, as shown in Fig. 2, the validation set is near to the training set in terms of prediction values. To make the evaluation more challenging, at each fold we set the subset as more than one set/fold away from the training set as the validation set. This can reduce the influence of redundant samples of neighbor sets on the over-estimation of the explorative prediction power.

To show how kmFCV can be used to achieve a stricter evaluation of the explorative power, we applied traditional CV, 1-step FCV (standard FCV), 2-step FCV, and 3-step FCV to evaluate RF and MLP for formation energy prediction problem. The fold value k for all methods is set as 100. The one-hot composition representation is used here to encode the materials in the MP formation energy dataset. The results are shown in Fig. 8 and Table 6.

First, from Fig. 8, we found that the prediction error MAEs increase with the increasing step m (the gap between the validation set and training set in terms of property values) for both RF and MLP. This is expected as the validation set becomes farther away from the training set, which makes it more difficult to predict the formation energy of those distant validation samples. This proves that m-step forward cross-validation is a more stringent method for evaluating explorative power of prediction models. In terms of exploration accuracy, the RF models have 0% for all m-step FCV while the MLP model increases with increasing m ranging from 20.52% to 36.29% (see Table 6). This is expected because the classification threshold stays the same as the largest property value in the training set no matter how the step value m changes. The further away from the validation set, the easier to classify validation samples into categories with lower or higher property values. These experiments demonstrate that neural networks are able to predict

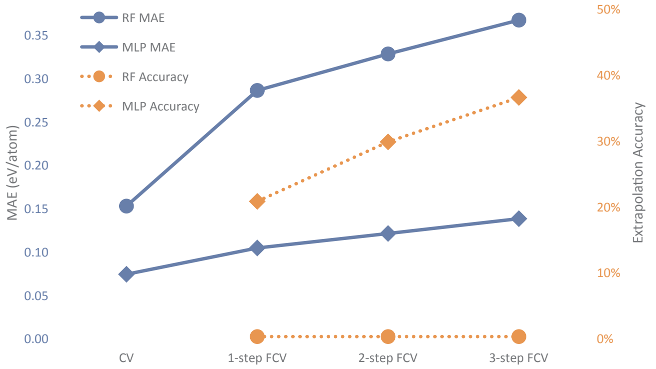


Fig. 8. Explorative power comparison of RF and MLP in terms of traditional CV, 1-step FCV, 2-step FCV, and 3-step FCV MAE and Exploration accuracy on formation energy predictions using one-hot composition representation.

Table 6
Explorative power comparison of RF and MLP in terms of traditional CV, 1-step FCV, 2-step FCV, and 3-step FCV MAE and Exploration accuracy on formation energy prediction using one-hot composition representation.

Benchmark problem	Model	Metric	CV	1-step FCV	2-step FCV	3-step FCV
Formation energy prediction	RF with one-hot composition representation	MAE (eV/atom)	0.1505	0.2839	0.3260	0.3650
		$E_{accuracy}$	N/A	**0%**	**0%**	0%
	MLP with one-hot composition representation	MAE (eV/atom)	0.0719	0.1022	0.1189	0.1360
		$E_{accuracy}$	N/A	20.52%	29.55%	36.29%

property values outside of the range of the training samples. It also shows that we can apply different step/gap size m to set up the level of difficulty in evaluating explorative prediction power of models.

5. Conclusions

We identified a special category of explorative materials property prediction problems in new materials discovery, which needs to predict property values out of the range of the training set. The limitations of traditional k-fold cross-validation methods for evaluating the explorative prediction performance of machine learning models for such problems are discussed. Accordingly, we proposed a family of k-fold-m-step forward cross-validation (kmFCV) methods as a new evaluation approach for benchmarking machine learning algorithms for explorative prediction. A new exploration accuracy metric is also proposed to directly reflect the explorative power. We applied forward CV and traditional CV to evaluate the performance of five machine learning models for three materials property prediction problems. Our comprehensive benchmark results showed that traditional CV methods tend to over-estimate the prediction performance when used for explorative materials property prediction. We also found 1NN and RF have no explorative power while MLP and CNN have better exploration capability, which however are still far from being satisfactory for discovering new materials with extremely low or high materials property values. Our work demonstrates the urgent need for improving the explorative power of the current machine learning models for new materials discovery.

CRediT authorship contribution statement

Zheng Xiong: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing - original draft, Writing - review & editing. **Yuxin Cui:** Investigation. **Zhonghao Liu:** Methodology. **Yong Zhao:** Methodology. **Ming Hu:** Investigation, Project administration, Writing - review & editing. **Jianjun Hu:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing - original draft, Writing - review & editing.

Data availability

All the source codes and benchmark datasets are freely available at https://github.com/buptxz/kmFCV.

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.

Acknowledgments

Research reported in this publication was supported in part by the NSF and SC EPSCoR/IDeA Program under award number (NSF Award #OIA-1655740 via SC EPSCoR/IDeA GEAR-CRP2019 19-GC02) and 1905775. The views, perspective, and content do not necessarily represent the official views of the SC EPSCoR/IDeA Program nor those of the NSF.

Author contributions

J.H. and Z.X. conceived the study; Z.X, J.H, and M.H. designed the research; Z.X., J.H, and Z.L. wrote the manuscript. Y.C and Z.L prepared figures. Z.X. and Y.Z prepared the dataset. J.H and M. H. supervised the study. All authors discussed the results and reviewed the manuscript.

References

[1] M.H. Esfe, et al., Optimization, modeling and accurate prediction of thermal conductivity and dynamic viscosity of stabilized ethylene glycol and water mixture Al2O3 nanofluids by NSGA-II using ANN, Int. Commun. Heat Mass Transfer 82 (2017) 154–160.

[2] Z.X. Yu, et al., Exceptionally high ionic conductivity in Na3P0.62As0.38S4 with improved moisture stability for solid-state sodium-ion batteries, Adv. Mater. 29 (2017).

[3] R. Bhattacharyya, S. Das, S. Omar, High ionic conductivity of Mg2+-doped non-stoichiometric sodium bismuth titanate, Acta Mater. 159 (2018) 8–15.

[4] V. Stanev, et al., Machine learning modeling of superconducting critical temperature, NPJ Comput. Mater. 4 (2018) 29.

[5] J. Turney, E. Landry, A. McGaughey, C. Amon, Predicting phonon properties and thermal conductivity from anharmonic lattice dynamics calculations and molecular dynamics simulations, Phys. Rev. B 79 (2009) 064301.

[6] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling, Phys. Rev. X 4 (2014) 011019.

[7] B. Meredig, et al., Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B 89 (2014) 094104.

[8] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, NPJ Comput. Mater. 3 (2017) 54.

[9] C. Kim, G. Pilania, R. Ramprasad, From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown, Chem. Mater. 28 (2016) 1304–1311.

[10] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, NPJ Comput. Mater. 2 (2016) 16028.

[11] L. Ward, et al., Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations, Phys. Rev. B 96 (2017) 024104.

[12] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (2018) 145301.

[13] X. Zheng, P. Zheng, R.-Z. Zhang, Machine learning material properties from the periodic table using convolutional neural networks, Chem. Sci. (2018).

[14] O. Isayev, et al., Universal fragment descriptors for predicting properties of inorganic crystals, Nat. Commun. 8 (2017) 15679.

[15] R. Liu et al., in: Proceedings of ACM SIGKDD Workshop on Large-scale Deep Learning for Data Mining (DL-KDD), pp. 1–7.

[16] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, Representation of compounds for machine-learning prediction of physical properties, Phys. Rev. B 95 (2017) 144110.

[17] K. Kim, et al., Machine-learning-accelerated high-throughput materials screening: discovery of novel quaternary Heusler compounds, Phys. Rev. Mater. 2 (2018) 123801.

[18] D. Jha, et al., Elemnet: deep learning the chemistry of materials from only elemental composition, Sci. Rep. UK 8 (2018) 17593.

[19] W. Kohn, Nobel Lecture: electronic structure of matter—wave functions and density functionals, Rev. Mod. Phys. 71 (1999) 1253.

[20] A. Jain, et al., Commentary: the Materials Project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 011002.

[21] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65 (2013) 1501–1509.

[22] S. Kirklin, et al., The Open Quantum Materials Database (OQMD): assessing the

accuracy of DFT formation energies, NPJ Comput. Mater. 1 (2015) 15010.

[23] S. Curtarolo, et al., AFLOWLIB. ORG: a distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58 (2012) 227–235.

[24] B. Meredig, et al., Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery, Mol. Syst. Des. Eng. 3 (2018) 819–825.

[25] H.S. Stein, D. Guevarra, P.F. Newhouse, E. Soedarmadji, J.M. Gregoire, Machine learning of optical properties of materials–predicting spectra from images and images from spectra, Chem. Sci. 10 (2019) 47–55.

[26] M. Schwarting, S. Siol, K. Talley, A. Zakutayev, C. Phillips, Automated algorithms for band gap analysis from optical absorption spectra, Mater. Discover 10 (2017) 43–52.

[27] B. Meredig, et al., Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery, Mol. Syst. Des. Eng. (2018).

[28] G. Martius, C.H. Lampert, Extrapolation and learning equations. arXiv preprint arXiv:1610.02995, 2016.

[29] S.S. Sahoo, C.H. Lampert, G. Martius, Learning Equations for Extrapolation and Control, arXiv preprint arXiv:1806.07259, 2018.

[30] M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, Science 324 (2009) 81–85.

[31] Science, N. I. o. M. SuperCon. (2011).

[32] A. Liaw, M. Wiener, Classification and regression by random forest, R news 2 (2002) 18–22.

[33] L. Ward, et al., Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.

[34] S.P. Ong, et al., Python materials genomics (pymatgen): a robust, open-source python library for materials analysis, Comput. Mater. Sci. 68 (2013) 314–319.

[35] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[36] D.P. Kingma, J. Ba Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

[37] Y. Lin, Y. Jeon, Random forests and adaptive nearest neighbors, J. Am. Stat. Assoc. 101 (2006) 578–590.

[38] Q. Zhou, et al., Learning atoms for materials discovery, Proc. Natl. Acad. Sci. 115 (2018) E6411–E6417.