Acta Materialia 240 (2022) 118353

Contents lists available at ScienceDirect

Acta Materialia

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

Full length article

On the value of popular crystallographic databases for machine learning prediction of space groups

Vishwesh Venkatraman[a,*], Patricia Almeida Carvalho[b,c]

[a] Norwegian University of Science and Technology, Trondheim 7491, Norway
[b] SINTEF Materials Physics, Oslo 0373, Norway
[c] CEFEMA, Instituto Superior Tecnico, University of Portugal, Lisbon, Portugal

ARTICLE INFO

Article history:
Received 25 February 2022
Revised 5 September 2022
Accepted 7 September 2022
Available online 10 September 2022

Keywords:
Space group
Machine learning
Multiclass
Multilabel
High entropy compounds

ABSTRACT

Predicting crystal structure information is a challenging problem in materials science that clearly benefits from artificial intelligence approaches. The leading strategies in machine learning are notoriously data-hungry and although a handful of large crystallographic databases are currently available, their predictive quality has never been assessed. In this article, we have employed composition-driven machine learning models, as well as deep learning, to predict space groups from well known experimental and theoretical databases. The results generated by comprehensive testing indicate that data-abundant repositories such as COD (Crystallography Open Database) and OQMD (Open Quantum Materials Database) do not provide the best models even for heavily populated space groups. Classification models trained on databases such as the Pearson Crystal Database and ICSD (Inorganic Crystal Structure Database), and to a lesser extent the Materials Project, generally outperform their data-richer counterparts due to more balanced distributions of the representative classes. Experimental validation with novel high entropy compounds was used to confirm the predictive value of the different databases and showcase the scope of the machine learning approaches employed.

1. Introduction

Crystal structure information is key to predicting materials performance. In particular, the underlying symmetry (i.e. crystallographic space group), which together with the atoms in the asymmetric unit and the cell constants define the structure adopted during crystallization. Demand for optimal properties has led to an increased focus on materials screening through both efficient experimental strategies [1] and computational methods, the latter being at the forefront of various data-driven initiatives [2], such as the Materials Project (MP) [3], JARVIS [4], AFLOW [5], and NOMAD [6] to name a few. On the experimental side, databases such as the Inorganic Crystal Structure Database [7] (ICSD) and the Crystallography Open Database [8] (COD) provide valuable crystal structure data.

Exhaustive exploration via experimentation is prohibitive and computer-aided materials discovery by way of both high-throughput *ab initio* simulations [9–11] and artificial intelligence [12–15] facilitates identification of interesting compositions in a timely manner. *Ab initio* determination of crystal structures poses however considerable challenges [16,17]. Indeed, although first-principles methods based on density functional theory (DFT) are quite popular for comparing the thermodynamic stability of structures, they require educated guesses on the candidate unit cells. In addition, they are computationally expensive and can be inaccurate, or even unreliable, when entropy effects are relevant but only ground state enthalpy is evaluated [18].

Recently, various data-efficient machine learning approaches to crystal structure prediction have been proposed that rival in accuracy with DFT methods [19–24]. Nonetheless, suitable representation of the materials, i.e. input to the machine learning models, is still required. The strategies adopted thus far to predict space group based on structure-derived characteristics [25–27] range from employing atomic pair distribution functions [25] to the use of string- and graph-based structural representations [28,29]. For instance, in the approach proposed by Liu et al [25], where a convolutional neural network predicts the space group of a structure given the corresponding atomic pair distribution function, the model achieved a top-6 accuracy of over 90% for 45 of

[*] Corresponding author.
E-mail address: vishwesh.venkatraman@ntnu.no (V. Venkatraman).

https://doi.org/10.1016/j.actamat.2022.118353
1359-6454/© 2022 The Author(s). Published by Elsevier Ltd on behalf of Acta Materialia Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

the most heavily represented space groups in ICSD. X-ray diffraction patterns of known crystal structures have also been used as structure-derived representations to predict space groups from non-interpreted diffraction data [30].

On the other hand, there has been a growing interest in structure-agnostic approaches, which do not require structural characteristics but instead learn directly from elemental compositions [26,31–34]. In these cases, sets of descriptors are constructed from compound stoichiometry and elemental properties assumed relevant [33,34] or, alternatively, are derived from composition through deep learning [31,35].

In practice there is an additional degree of complexity regardless of the strategy employed. Each composition often corresponds to multiple crystal structures, with the one adopted depending on the specific crystallization conditions (pressure, temperature and kinetic factors). Zhao et al attempted to address this problem through multilabel classifiers for polymorphism prediction by assigning to each sample a set of target labels [27]. The results obtained using the MP database yielded F1-scores (weighted average of the precision and recall) of 0.65 for the multiclass prediction of space groups.

Machine learning studies for space group prediction typically use either ICSD *or* MP data. In this work we have compared five different databases popular in the materials science community, three of which contain experimental data, namely COD [8], Pearson's Crystal Data (PEARSON) [36] and ICSD [7], with the other two comprising DFT-calculated structures, namely the Open Quantum Materials Database (OQMD) [37] and MP [3]. We have used structure-agnostic models trained on fixed-length descriptors derived from the chemical formulas, while polymorphism was explored through multilabel classifiers. In addition, a deep-learning strategy producing a stoichiometry-to-descriptor map directly from the data [31] was employed for comparison. In all cases, the performance was evaluated in terms of the ability to classify any given composition into a specific space group. However, since many of the 230 space groups are sparsely populated, we have adopted the same strategy as Liu et al. [25] and performed the analysis solely for the 50 most frequent classes in each database. Furthermore, we have independently tested the predictive value of the databases for less stringent classifications into 7 crystal systems, 5 lattice centering options, 14 Bravais lattices, and 32 point groups. Experimenting with multiple datasets using different machine learning approaches and multiple end-points has not been carried out before to the best of the authors knowledge.

Predicting structural information is particularly relevant in the field of high-entropy (HE) materials. HE alloys were originally defined as equimolar solid solutions of five or more metals for which high configurational entropy inhibits long-range order [38,39]. In fact, entropy alone is not a good predictor of full solubility and, like any process governed by the total free energy, ensuring stability requires supplementation with a low-enthalpy criterion [40]. Nonetheless, the concept of maximizing configurational entropy to enhance solid-state miscibility continues to inspire the exploration of unfamiliar composition spaces and the popular, albeit imprecise, high-entropy designation seems destined to endure [40]. The spotlight has been on the mechanical properties of HE alloys, but interest in functional behaviour is swiftly rising. In particular, the vast potential of combining metal solid solutions with structural boron, carbon, oxygen or silicon to form HE compounds is becoming evident [41], but *a priori* knowledge on the crystal structure adopted for a given composition remains crucial. The structure-agnostic models trained on the different databases have been applied to the particular case of HE borides, carbides, oxides, silicides and antimonides with experimentally validated crystal structures.

2. Methods and materials

2.1. Datasets

Five different databases have been analyzed: (i) COD [8,42,43] (ii) PEARSON [36], (iii) OQMD [37,44] (iv) MP [3] and (v) ICSD [7]. COD, ICSD and PEARSON comprehend curated collections of experimentally solved crystal structures, while OQMD and MP offer computed information, namely DFT-calculated thermodynamic and structural properties, for known and predicted materials. Among these, COD, OQMD and MP are open access, while ICSD and PEARSON are commercial databases.

For each dataset, we eliminated duplicate entries. Compounds for which the formulas could not be parsed and structures that contained only a single element or noble gas(es) were excluded, as well as those that could not be mapped to the 230 crystallographic space groups. After cleansing, the available data were divided into two categories for which (i) there was a one–to–one association between composition and space group or (ii) the composition belonged to multiple space groups. Table 1 provides a summary of these two categories in the processed data together with the fraction of compounds containing oxygen, a distinctive characteristic of the datasets. The first category was used for the multiclass studies and the second one for the multilabel studies.

The heatmap in Fig. 1 shows the distribution of the 230 space groups across the 5 datasets for the one–to–one category. The data in COD cover all space groups, while a number of space groups are missing in the other databases (1 in ICSD, 3 in MP, 25 in OQMD and 6 in PEARSON). $P\bar{1}$, $P2_1/c$, $C2/m$, $Pnma$ and $Fm\bar{3}m$ are among the most frequent space groups in all databases.

The pie charts in Fig. 2a show that the databases exhibit different crystal system distributions. COD contains large proportions of monoclinic, triclinic and orthorhombic structures, reflecting a high fraction of low-symmetry mineral compounds. ICSD, PEARSON and MP exhibit relatively even distributions of the 7 crystal groups, while OQMD is dominated by cubic structures, which probably stems from a practical interest in high-symmetry compounds relevant for mechanical, optical and electronic applications. Fig. 2b shows the distribution of the 5 lattice centering types. Primitive cells largely prevail in COD; ICSD, Pearson and MP show lower prevalence of primitive cells and comparable distributions of all centering options; while face centered cells are notably preponderant in OQMD.

The heatmap in Fig. 3a shows the distribution of Bravais lattices, which reflects the combined frequency of crystal system and cell centering options. The heatmap in Fig. 3 b shows the distribution of the 32 point groups across the 5 datasets, where the prevalent groups in COD ($\bar{1}$ and $2/m$) and OQMD ($\bar{4}3m$ and $m\bar{3}m$) stand out.

The compounds in the one–to–one category (see Table 1) were further analyzed with respect to the number of elements. Fig. F1 in the Supplementary Material shows the distribution of crystal systems in each subset of multi-element compounds in the different databases. In all cases, only a residual number of compounds consists of more than 9 elements (see also Table S6 and Figs. F2– F4 in the Supplementary Material).

A number of compounds in the different datasets exhibit one–to–many relationships. However, while OQMD has less than 20 compounds in these circumstances, the other four databases contain much larger numbers (see Table 1). As shown in Fig. 4, compounds belonging to 2 space groups are the predominant class in the context of polymorphism in all databases, while only minor fractions were associated with more than 4 space groups.

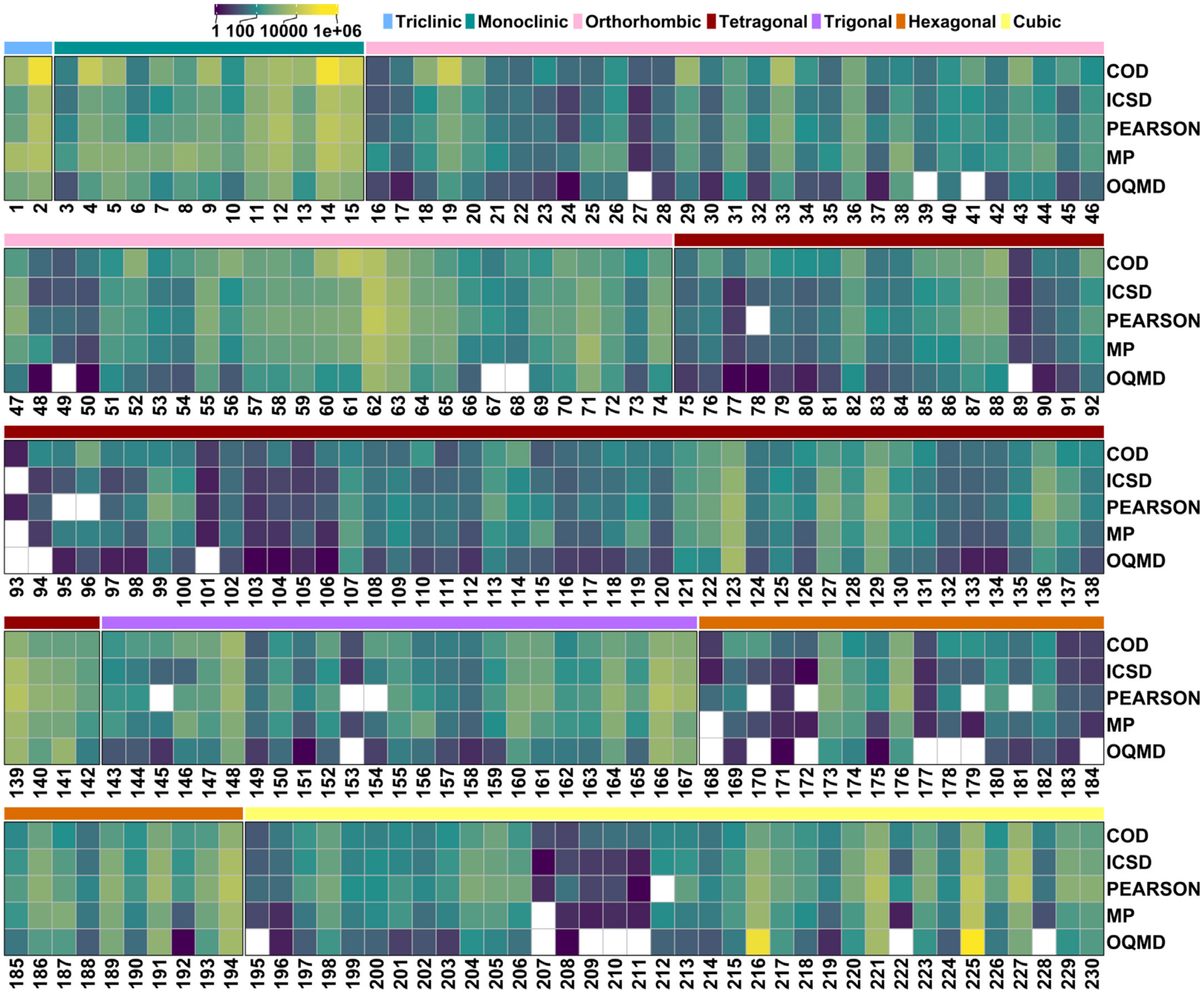


Fig. 1. Heatmap showing the distribution of the 230 space groups in COD, ICSD, PEARSON, MP and OQMD. See also Table S1 in the Supplementary Material.

Table 1
Number of entries in each dataset in terms of compound mapping to 1 space group (one–to–one) or more (one–to–many). The last column lists the percentage of oxygen-containing compounds.

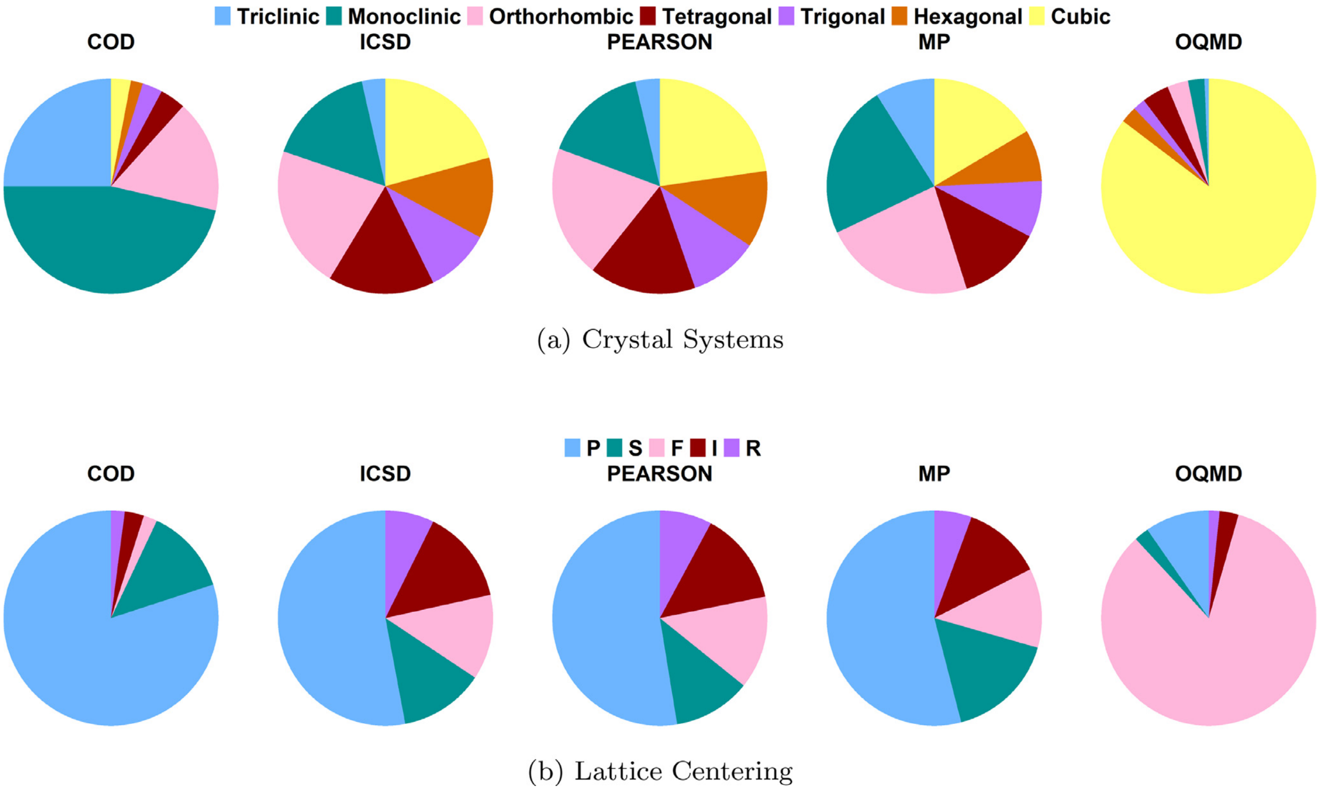
Dataset	#1:1	#1:many	% O-containing compounds
COD	329,080	24,016	73
ICSD	99,646	7327	48
PEARSON	160,440	13,489	53
MP	83,781	14,362	45
OQMD	322,338	-	6

2.2. Compound representation

For each compound, a set of over 200 descriptors was constructed from the respective composition and elemental properties, such as electronegativity [45], atomic weight, Zunger pseudopotential radius [46], Mendeleev number, polarizability, heat of formation, number of filled/unfilled valence orbitals, heat of fusion etc., which were extracted from the python package *Mendeleev* [47] and from http://www.knowledgedoor.com. The fixed-length descriptors were based on maximum, minimum, fraction-weighted mean and mode, as well as on the average deviations of the elemental properties compared to the ones of the prevalent element [48,49]. The descriptors were calculated using software developed in-house (available from https://github.com/vvishwesh/MaterialDescriptors), which extends the Magpie [50] set to include additional elemental attributes. The list of the descriptors is provided in the Supplementary Material.

2.3. Modelling and assessment

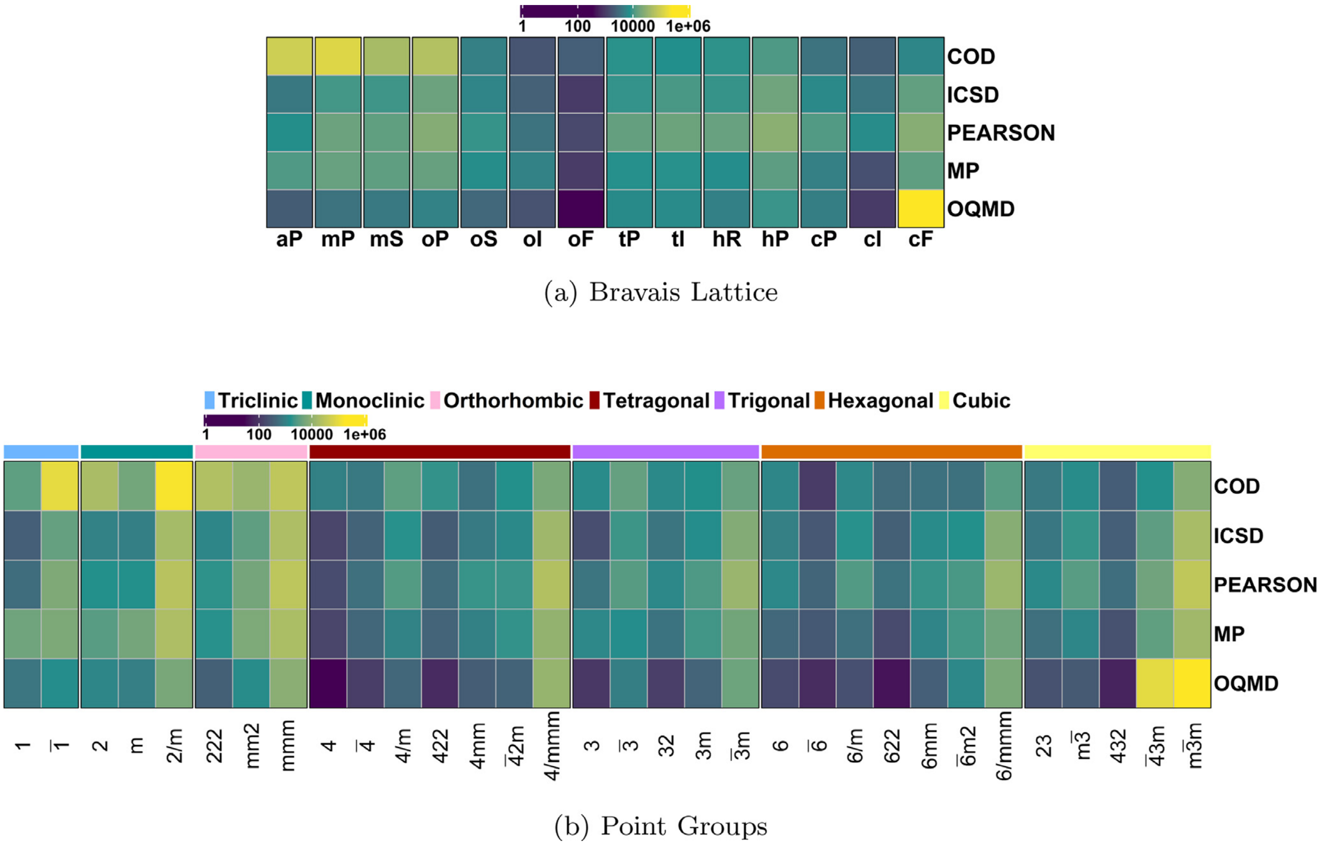
Classification models were built using random forests [51] employing the *ranger* [52] machine learning package available in the R Project for Statistical Computing [53]. Random forests uses an ensemble of decision trees wherein predictions from multiple tree models are combined. The algorithm is easy to train and has been shown to provide robust models in a number of modelling tasks spanning multiple fields from materials property prediction [48,54], drug design [55], imaging [56] to precision medicine [57]. In the random forest model, the number of decision trees was set to 500 while the number of variables randomly sampled as candidates at each split was varied between 2 and the number of input variables. Training and validation of the models was carried out



(a) Crystal Systems

(b) Lattice Centering

Fig. 2. Distribution of the (a) 7 crystal systems and (b) 5 lattice centering types (P – Primitive, S – Base-centered, I – Body-centered, F – Face-centered and R – Rhombohedral. See also Tables S2 and S3 in the Supplementary Material.



(a) Bravais Lattice

(b) Point Groups

Fig. 3. Heatmaps showing the distribution of the (a) 14 Bravais lattices; (a – Triclinic, m – Monoclinic, o – Orthorhombic, t – Tetragonal, h – Hexagonal and c – Cubic, P Primitive, S – Base-centered, I – Body-centered, F – Face-centered and R – Rhombohedral), and (b) 32 point groups. See also Tables S4– S5 in the Supplementary Material.

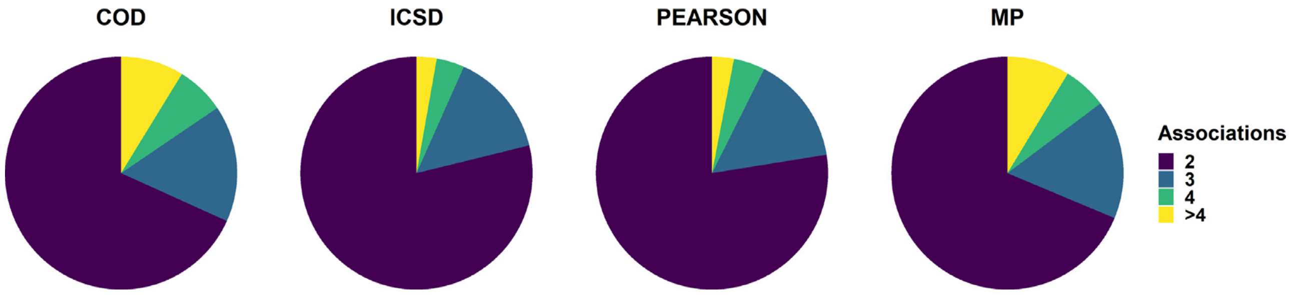


Fig. 4. Distribution of one–to–many associations (compounds mapping to more than 1 space group). The total number of one–to–many associations found in each dataset is listed in Table 1. Few compounds showed 1–many associations in OQMD and this database is omitted from the analysis.

using independent splits of the data into calibration (80%) and test sets (20%). The calculations were repeated 5 times to check for any large deviations in performance caused by the random partitioning. In order to reduce the dimensionality of the feature space, a pairwise squared correlation cut-off of 0.90 was applied to the training set. The feature removal resulted in a reduction of the number of variables to 85–140 depending on the particular dataset. Since many of the 230 space groups were sparsely populated, we have performed the analysis solely on the 50 most frequent classes in each dataset. In addition, we developed separate models for the less stringent classifications into 7 crystal systems, 5 lattice centering options, 14 Bravais lattices, and 32 point groups, which taken together define the space groups.

Only space groups and crystal systems were considered for the multilabel classification, which involved assigning simultaneously multiple labels (polymorphs) to a composition. The multilabel classification was carried out using the multivariate random forest algorithm[58] available in the *mlr*[59,60] package included in R. For the multilabel space group prediction, only the 10 most frequently occurring space groups in each one–to–many datasets were considered. Due to insufficient data, OQMD was excluded from the multilabel analysis.

Model performance was assessed using 5-fold cross-validation. Response randomization (*y*-randomization) was additionally carried out to ensure that the achieved performance has not resulted from chance. For all classification problems, evaluation metrics that are sensitive to class imbalance [61,62] have been used. In addition to the standard accuracy, the other performance metrics employed include:

$$Precision = \frac{1}{m}\sum_{i}^{m}\frac{tp_i}{tp_i + fp_i}$$

$$Recall = \frac{1}{m}\sum_{i}^{m}\frac{tp_i}{tp_i + fn_i}$$

$$F1 - score = \frac{1}{m}\sum_{i}^{m}\frac{2 * (Recall_i \times Precision_i)}{(Recall_i + Precision_i)}$$

where for class C_i (there being m such classes), tp_i are the true positive, and fp_i – false positive, fn_i – false negative, and tn_i – true negative counts, respectively. For multiclass classification, the macro-averaged F1-score was used, i.e. the F1 scores were computed for each class and then averaged via arithmetic mean, thereby treating all classes equally.

2.4. Experimental methods

In the context of screening new materials a set of 7 new transition-metal silicides with high configurational entropy and unknown space groups have been produced with an Arc Melter (Buhler, Model AM500). The composition and symmetry of the new phases were determined, respectively, by X-ray energy dispersive spectroscopy (EDS) and electron backscattered diffraction (EBSD) with an FEI NanoLab 600 instrument equipped with Oxford Instruments EDS and EBSD detectors. In addition, several high entropy compounds recently reported in the literature (see Table S10 in the Supplementary Material for a full list of the compounds and associated references), have been included in the set of compositions used to interrogate the databases.

3. Results and discussion

3.1. Multiclass classification

Multiclass models for the prediction of space groups and separately for crystal systems, lattice centerings, Bravais lattice and point groups have been investigated. Although these other categories are embedded in the space group designation, they have been treated as independent entities to assess differences in performance as a function of dataset size and class size uniformity. This strategy also allows to retrieve reliable structural information for compounds belonging to least populated space groups (excluded from the space group models).

A summary of model performance for the 5 types of multiclass evaluation carried out using the different datasets is provided in Fig. 5. Detailed performance statistics are shown in Table S7 in the Supplementary Material. In all cases, the statistics obtained for the test sets mirror those of the training sets. For 5 independent iterations, where the models were trained on multiple random splits of the data, standard deviations for the calculated statistics were $\approx \pm 0.03$, which suggests that data splitting has not impacted performance.

In the case of crystal system, the models trained on PEARSON and ICSD yielded relatively high F1-scores ($\gtrsim 0.70$), while a considerably lower value was achieved with COD ($F1 \approx 0.50$). The performance obtained for the test set (see heatmaps of the confusion matrices in Fig. F5 in the Supplementary Material) indicates that the COD model shows high predictive ability for the prevalent monoclinic system, but that the performance for the other crystal systems is much lower due to the data skewness (see Fig. 2a). A similar trend is seen for the model trained on OQMD, which shows the best predictive ability for the overwhelmingly dominant cubic system. However, here the effect of skewness is less pronounced, with the lower performance for non-cubic systems resulting from the inability to discriminate between these less frequently occurring classes (see Fig. F5 in the Supplementary Material). In contrast, the models trained on ICSD and PEARSON exhibit average to good performance for all crystal systems, while the predictive value of MP is more modest and on average comparable to that of OQMD.

The predictions for Bravais lattice, point group, lattice centering and space group followed trends similar to those observed for crystal system: the models trained on PEARSON and ICSD outper-

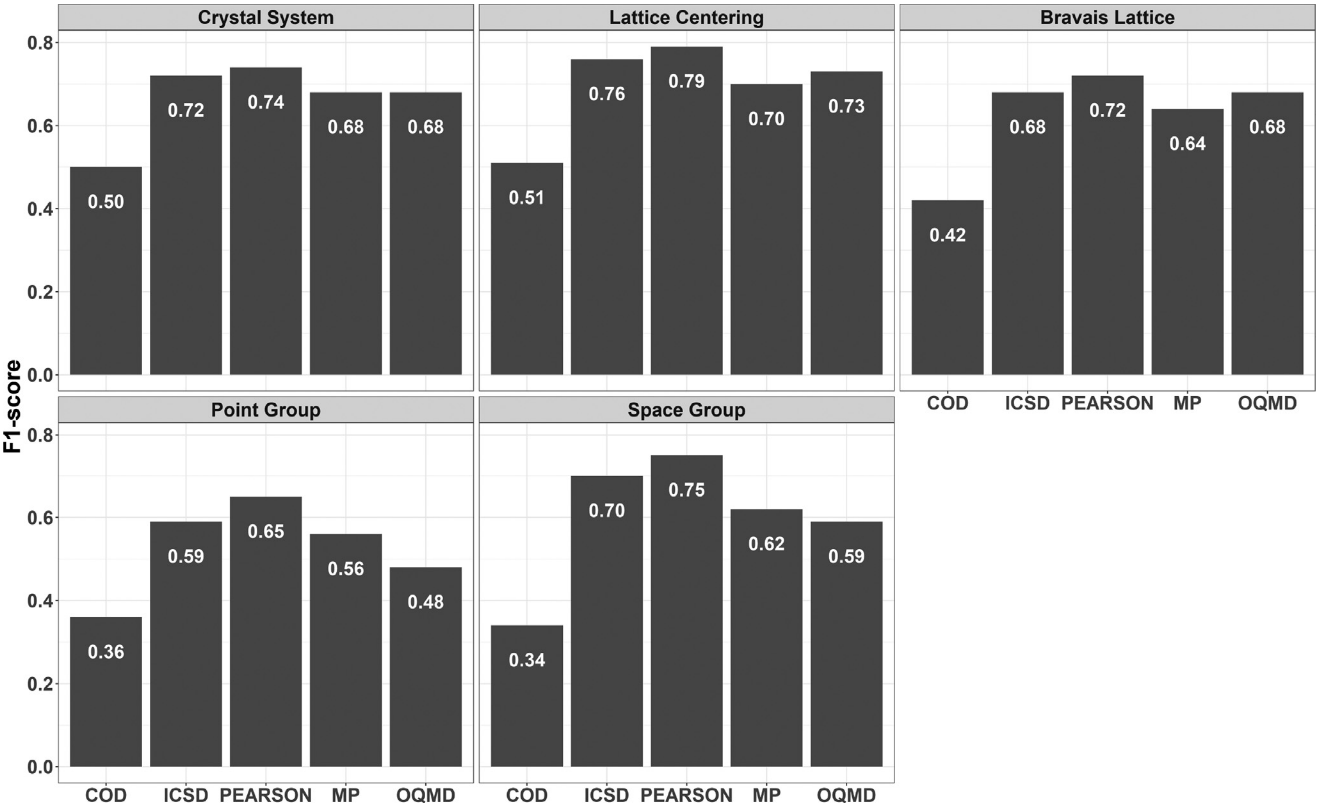


Fig. 5. Performance of the random forests models in terms of the F1-scores obtained for the validation sets presented in the typical crystallography order, i.e., crystal systems (7 classes), lattice centering options (5 classes), Bravais lattices (14 classes), point groups (32 classes) and space groups (50 most populated of the 230 classes). Values shown are an average of 5 independent runs. For detailed performance statistics see Table S7 in the Supplementary Material.

formed the others, while the opposite occured for models trained on COD (see Fig. F6 in the supplementary material for the space group results). In general, performance decreased with increasing number of classes (Fig. 5). However, in the case of space group the performance was boosted by excluding the poorly populated classes from the dataset, making the prediction comparable to that achieved for lower class numbers (see Fig. F7 in the Supplementary Material). Nevertheless, this effect was marginal for COD due to its strong class imbalance. For example, among the 50 top populated space groups in COD there are more than 90,000 entries for $P\bar{1}$ but only ≈ 550 for $R3m$. Attempts to address class imbalance using methods such as undersampling/oversampling of, respectively, majority/minority classes [63] as well as other alternative learning algorithms such as XGBoost [64–66] did not improve the prediction performance (Table S11 in the Supplementary Material lists the F1 scores obtained).

Close examination of the test set predictions for space group allows to refine the discussion. The heatmap of the confusion matrix for the PEARSON model (Fig. 6) shows that most of the 50 space groups are well predicted and a similar behaviour is observed for ICSD (see Fig. F6 in the Supplementary Material). The confusion matrices for the other datasets show higher levels of misclassification for underrepresented classes, justifying the darker shades at the diagonals. Namely, for the COD model (Fig. 7), only a small set of the space groups (numbers 123, 129, 139, 160, 191, 194, 221, 225 and 227), reasonably well represented in the dataset, result in predictions with accuracy values close to 80%. Thus, in general, improved results are seen for datasets with more even class distributions. Relatively high misclassification occurred toward space groups 14 and 62. These errors are largely associated with the over-representation of these two classes in COD (see Fig. 1), in addition to a prevalence of oxygen-containing compounds in the low symmetry compounds, which form a significant majority in the experimental databases (similar observations were reported by Liang et al. [26]). Nevertheless, the prediction results showed that the second highest probability was typically associated with the target space group. As shown in Table S8 in the Supplementary Material, a performance boost in terms of the top-n accuracies is seen as n varies from top-1 to the top-3, i.e. for the majority of the models the correct answer was found to be among the three most probable predictions.

3.2. Fixed-length vs deep learning representation

Prediction models based on constructed descriptors (derived from stoichiometry) promote an understandable causality between inputs and outputs. In an alternative approach, Goodall and Lee [31] have recently proposed a deep learning framework called Roost (Representation Learning from Stoichiometry) that makes use of a message-passing neural network to directly learn material descriptors. Each compound corresponds to a dense weighted graph, where the nodes represent the different elements weighted by the corresponding molar fractions. Code from the Roost repository was used to generate deep learning predictions for comparison with the descriptor-based approach used above. Each database was randomly divided into training (60%), validation (20%) and test (20%) sets (see Table S12 for information on the hyperparameters). Table 2 summarizes the prediction performance of the deep learning models for both crystal systems and the most frequent 50 space groups in the 5 datasets. Except for OQMD, the performances of the deep learning models are lower than those obtained by the descriptor-based random forests approach. Indeed, although Goodall and Lee [31] have demonstrated good predictive ability of their deep learning method in regression problems tested

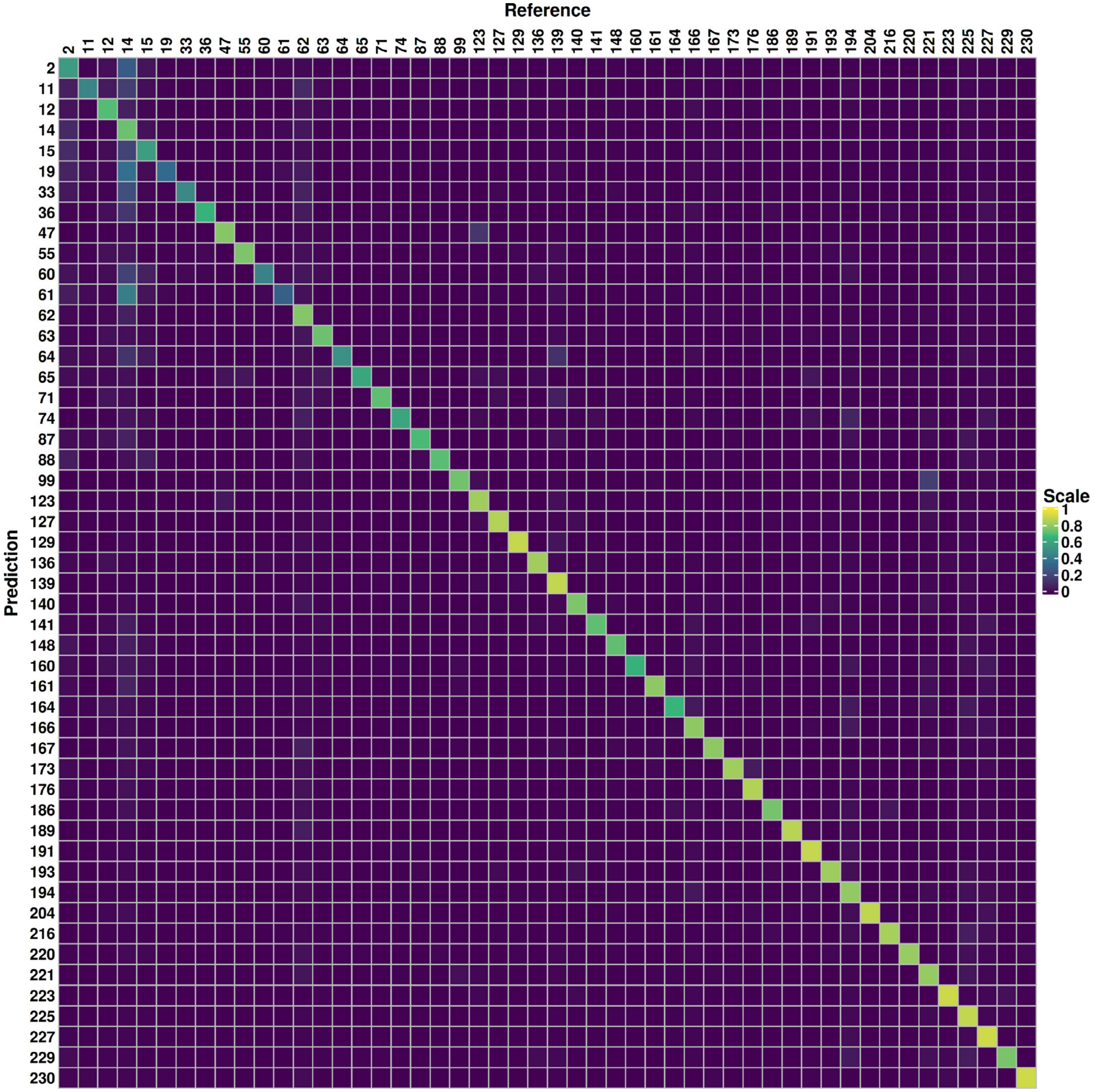


Fig. 6. Confusion matrix of the PEARSON model predicting the 50 most frequent space groups in the PEARSON database. Elements on the diagonal indicate the proportion of correct predictions (the integrated intensity across each row is 1.0 which ideally should be concentrated at the diagonal element). Off-diagonal brighter elements indicate incorrect classifications.

Table 2
Test set performance of the deep learning classification model for both crystal system and space group prediction (averaged through the 7 and 50 classes, respectively, for each dataset).

Source	Crystal System		Space Group	
	Accuracy	F1-score	Accuracy	F1-score
COD	0.47	0.39	0.34	0.25
ICSD	0.66	0.62	0.63	0.59
PEARSON	0.44	0.40	0.67	0.65
MP	0.60	0.58	0.54	0.50
OQMD	0.93	0.64	0.93	0.54

on OQMD, training on the other datasets generated rather disappointing results. Nonetheless, improvements in performance may be possible through suitable modifications of both the loss function and network architecture [31]. The addition a self-attention mechanism, which allows for learning of inter-element interactions, may also improve the prediction performance [35].

3.3. *Multilabel prediction*

The models for crystal system prediction in the polymorphism context (see Table 3) yielded F1-scores ranging from 0.54 (for ICSD) to 0.69 (for COD). The numbers compare favourably with

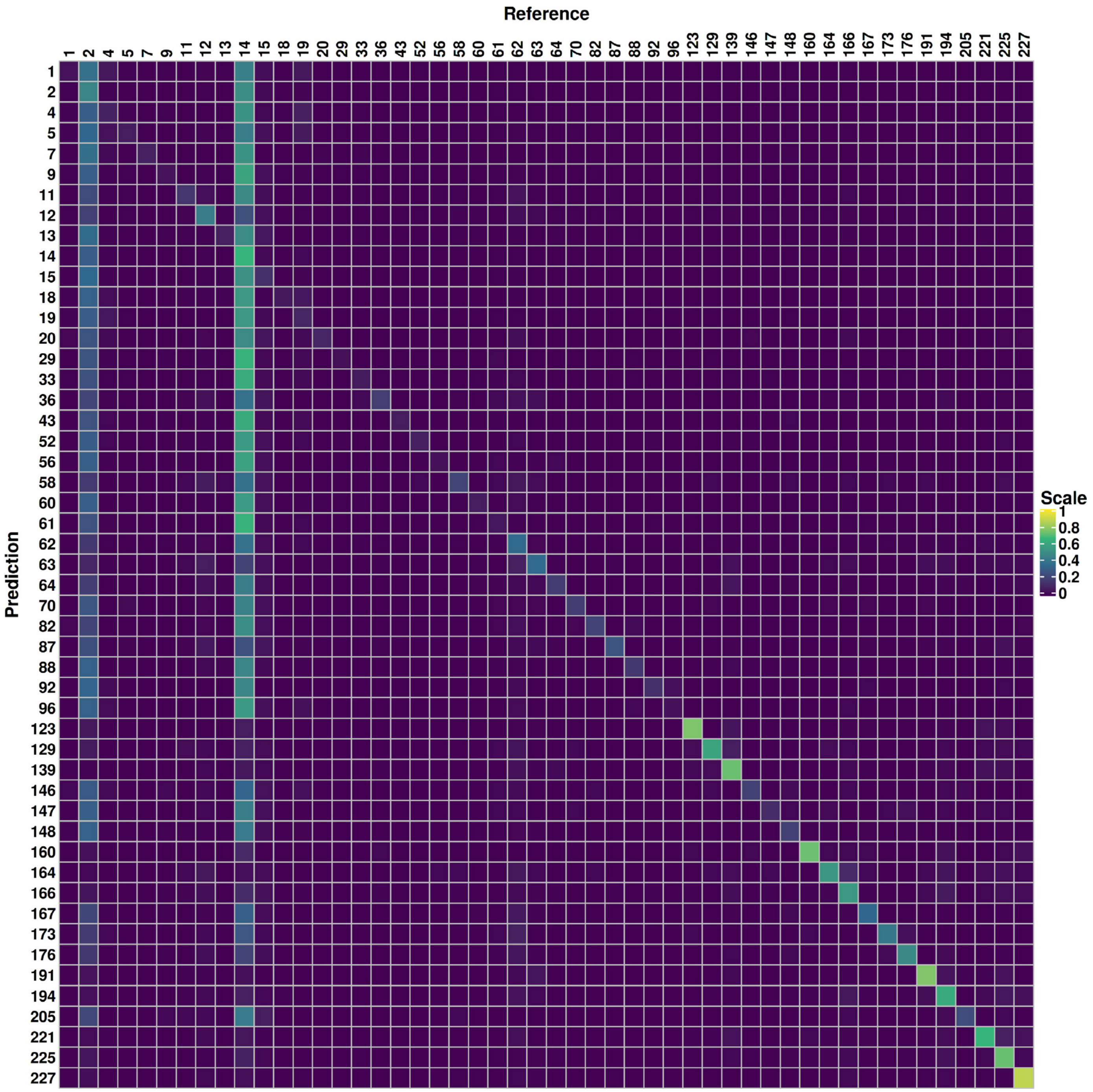
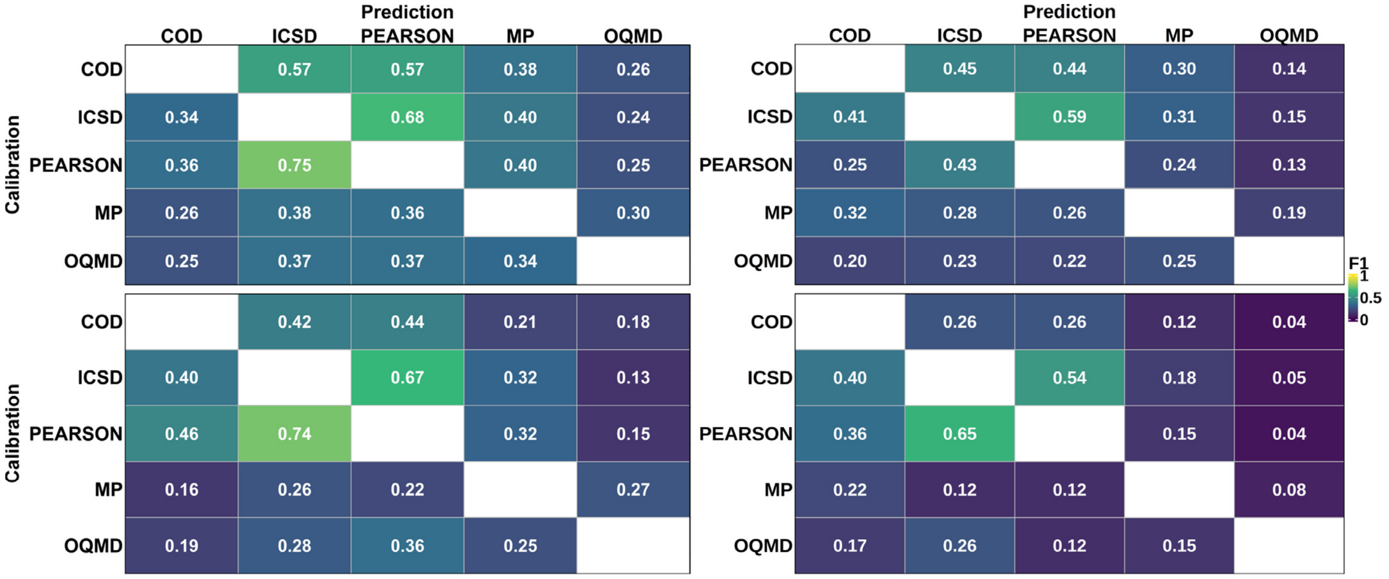


Fig. 7. Confusion matrix of the COD model predicting the 50 most frequent space groups in COD. Elements on the diagonal indicate the proportion of correct predictions (the integrated intensity across each row is 1.0 which ideally should be concentrated at the diagonal element). Off-diagonal brighter elements indicate incorrect classifications.

Table 3
Performance of the random forest-based multilabel classification model predictions of the 7 crystal groups (see Section 3.3 for additional details).

Source	Set	Accuracy	Precision	Recall	F1-score
COD	TRAIN	0.59	0.74	0.72	0.69
	TEST	0.59	0.74	0.73	0.69
ICSD	TRAIN	0.49	0.75	0.52	0.55
	TEST	0.48	0.76	0.52	0.54
PEARSON	TRAIN	0.55	0.79	0.59	0.61
	TEST	0.55	0.79	0.60	0.61
MP	TRAIN	0.55	0.75	0.62	0.63
	TEST	0.55	0.74	0.61	0.62

results reported by Zhao et al., who obtained similar F1-scores for descriptor-based models trained on MP data [27]. These results show that for polymorph crystal system prediction, COD outperforms the other databases, as expected from the larger size of its multilabel classes (see Fig. 4). For polymorph space group prediction, the models exhibited lower F1-scores with only COD nearing 0.50 (see Table 4). Given today's paucity of data and the large number of space group classes, building reliable multilabel models for space group prediction clearly requires investing in further populating the available crystallographic databases. On the machine learning front, Alsaui et al. [67], recently investigated prominent resampling techniques and have advocated the use of random un-



(a) RF

(b) DL

Fig. 8. Heatmaps of the F1-scores obtained for crystal system (top row matrices) and space group (bottom row matrices) predictions. Both random forests (RF) and deep learning (DL) models were trained independently on COD, ICSD, PEARSON, MP and OQMD, and then used to predict the other datasets.

Table 4
Performance of the random forest based multilabel classification model for the top 10 frequently occurring space groups in each dataset (see Section 3.3 for additional details).

Source	Set	Accuracy	Precision	Recall	F1-score
COD	TRAIN	0.43	0.65	0.48	0.52
	TEST	0.43	0.65	0.48	0.52
ICSD	TRAIN	0.48	0.78	0.27	0.49
	TEST	0.49	0.79	0.28	0.50
PEARSON	TRAIN	0.54	0.81	0.39	0.55
	TEST	0.54	0.82	0.39	0.56
MP	TRAIN	0.49	0.75	0.37	0.52
	TEST	0.49	0.75	0.37	0.52

dersampling (wherein randomly selected samples from the majority class are omitted) for improved classification performance in multiclass multilabel problems.

3.4. Cross-predictive ability

More challenging than predicting responses included in the datasets on which the models were trained is the question: How does performance vary when models trained on one dataset are used to predict data included in other datasets? Cross-predictive capabilities were assessed after excluding compounds common to the training sets to prevent biasing i.e. compositions that exist in the training or validation datasets, were removed from the test set. The results (in terms of F1-scores) for crystal system and space group are summarized in Fig. 8. The models trained on OQMD and MP show relatively low F1-scores and follow rather similar trends for both crystal systems and space groups. For crystal systems, the best overall cross-prediction performance (based on value averaging) was achieved by COD ($\bar{F}1 = 0.45$) followed by PEARSON ($\bar{F}1 = 0.44$). In the case of space groups, the best results were obtained using models trained on PEARSON and ICSD. Compared with the random forests model, the performance of the stoichiometry-based deep learning predictions was generally poorer, with the models trained on COD and ICSD yielding mean F1 values of 0.37 and 0.33 respectively for crystal system prediction. For space group prediction, the deep learning models followed similar trends to that of the random forests approach, albeit with lower F1-scores.

Models for crystal system prediction trained on ICSD and PEARSON when applied to each other yielded good F1-scores, with the similar class distributions contributing to the higher performance (see Fig. 2a and Figs. F8–F17 in the Supplementary Material). Contrarily, and despite the abundance of data in OQMD and COD, when models trained on either dataset are applied to the other the results are relatively poor since OQMD is dominated by compounds with cubic symmetry while low-symmetry crystal systems are prevalent in COD (Fig. 2a). In the case of space groups, the models trained on ICSD and PEARSON achieved good F1-scores for most of the 50 space groups used for training (see Figs. F9 and F12 in the Supplementary Material). While arguments based on similar ratios of compounds to classes are valid in the case of space groups, it must be pointed out that the 50 most frequent classes may not be the same in all databases and may differ strongly in the relative number of entries. In many cases, we observe that the models record a high precision but low recall as well as the converse (low precision and high recall) which implies that the model at times has a high false positive rate or a high false negative rate (predicted labels are incorrect) respectively. In particular, for space group the scores are impacted by the fact that some classes contain very few compounds (< 10). Examination of the top-n accuracy values, as seen in Fig. 9 (see also Table S9 in the Supplementary Material), shows that for a majority of the models, the top-2 accuracy values are much higher than those for the top-1 indicating that the models are unable to correctly select between the two top predictions.

The error rate variation showed diverse patterns with respect to the type of the training vs predicted database (experimental or theoretical) as well as to the number of elements in the compounds. For instance, for the model trained on COD, the error rates associated with space group prediction for ICSD and PEARSON data are below 45% and decrease with the number of elements in the compound, while for theoretical OQMD and MP data the error rates are $\geq 70\%$. Analysis of the errors also showed that training on datasets with relatively large proportion of O-containing compounds (COD – 75%, Pearson – 53%, ICSD – 48%, see Fig. 1)

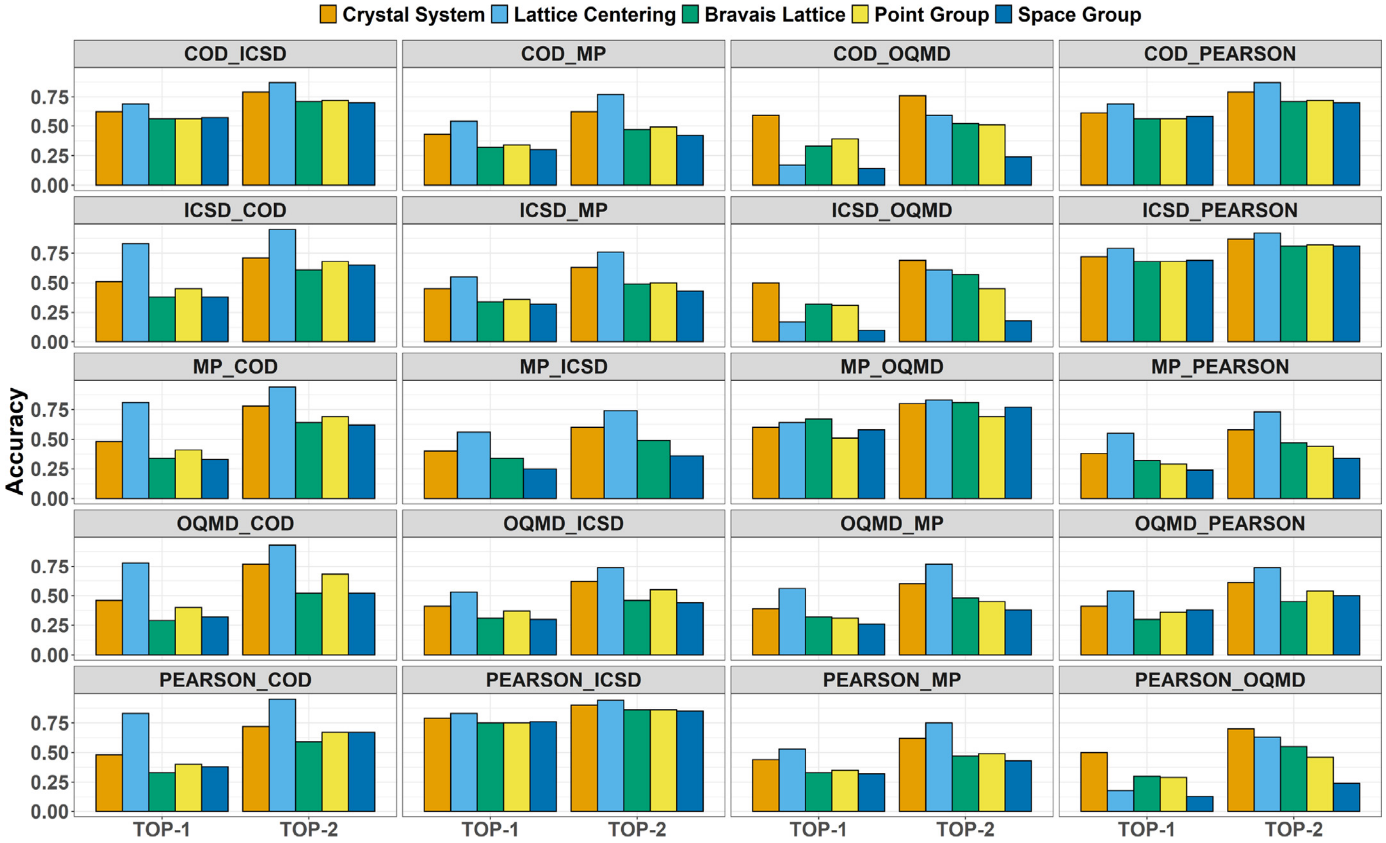


Fig. 9. Top-1 and top-2 accuracy of the random forest models when predicting (i) crystal system, (ii) lattice centering, (iii) Bravais lattice, (iv) point group and (v) space group. The accuracy values for X_Y correspond to the cross-prediction performances where a model trained on dataset X is applied to dataset Y.

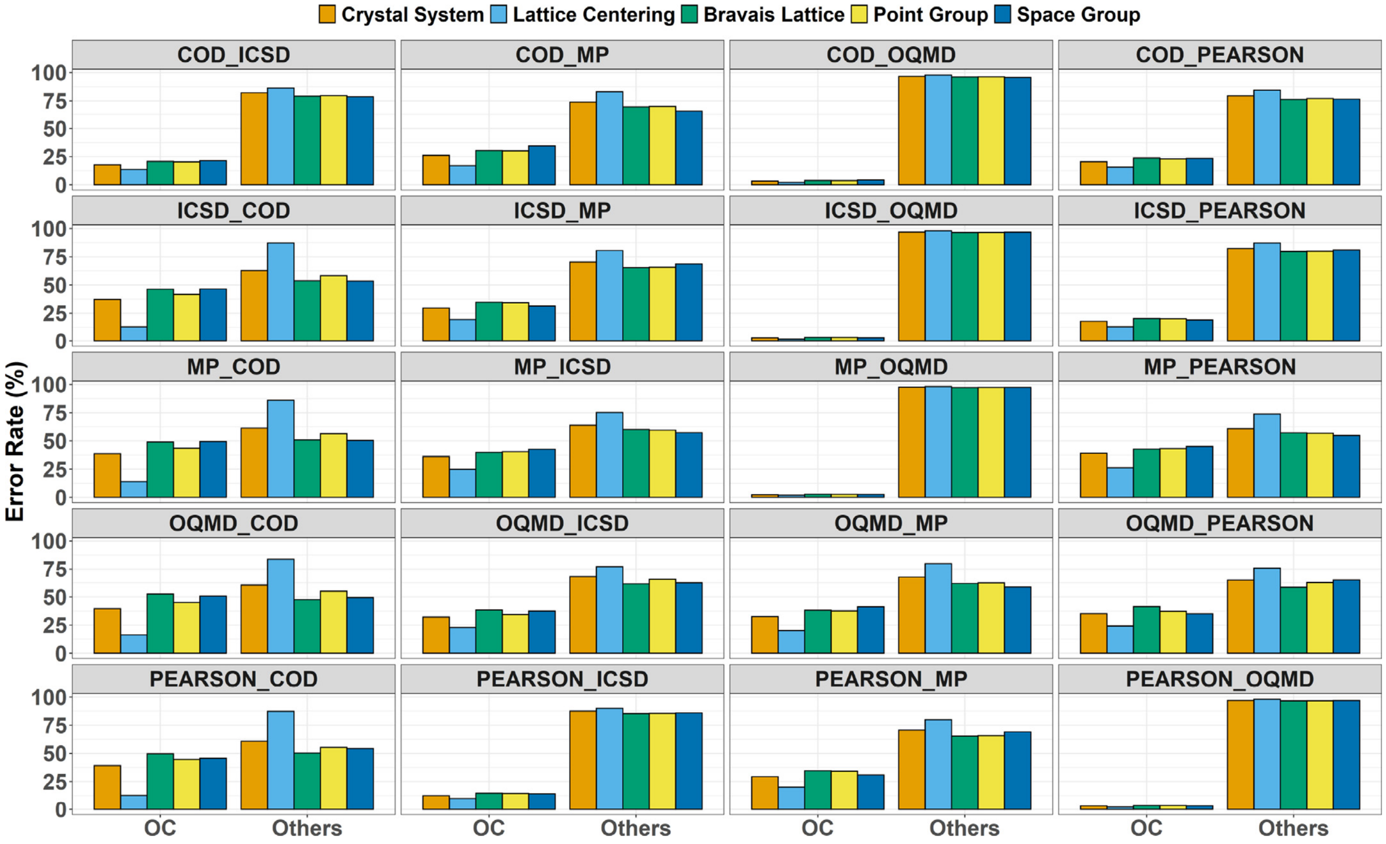


Fig. 10. Error rates of the random forest models when predicting (i) crystal system, (ii) lattice centering, (iii) Bravais lattice, (iv) point group and (v) space group for O-containing vs other compounds. The X_Y errors correspond to cross-prediction performances where a model trained on dataset X is applied to dataset Y.

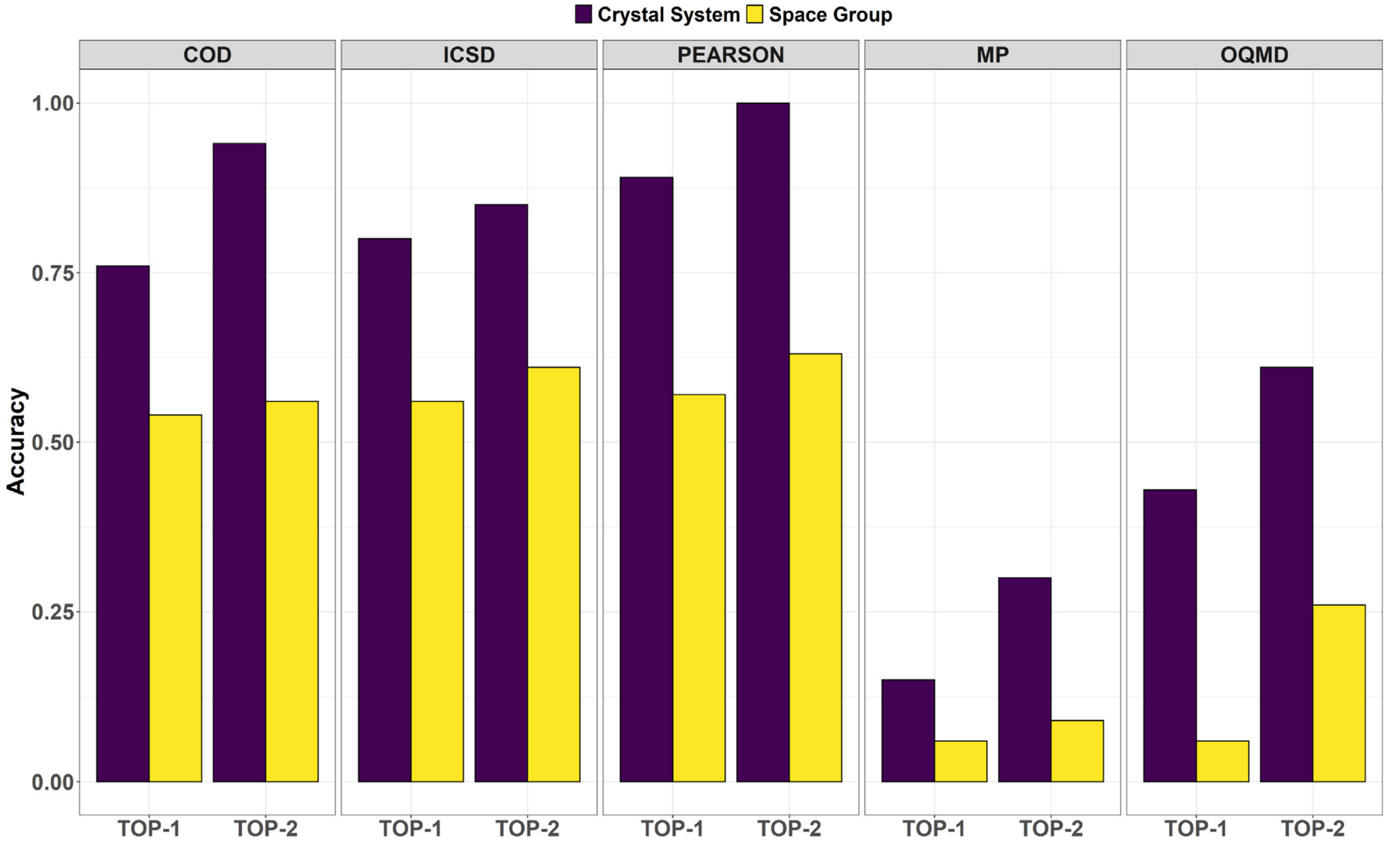


Fig. 11. Comparison of the ability of the random forest models calculated for the 5 datasets to predict crystal systems and space groups for a set of 54 high entropy compounds with experimentally verified space groups (see Section 2.4).

resulted in considerably lower error rates for predicting structural information of other O-containing compounds (see Fig. 10). This bias partly justifies the large errors observed when the models trained on COD, ICSD and Pearson are used to predict OQMD data (6% O-containing compounds).

The models were also tested on a set of 54 high entropy compounds with experimentally verified space groups. The performance for crystal system and space group predictions is summarized in Fig. 11. In the case of crystal system, the best performance was achieved with the model trained on PEARSON (89%), while the models trained on ICSD and COD attained slightly lower accuracies of 81% and 76%, respectively. For space groups, the model trained on PEARSON achieved much lower accuracy (57%), while other models performed rather poorly, with training on OQMD or MP leading to correct classification of only 3 compounds. The relatively low number of compounds with more than 4 elements in OQMD and MP may also contribute to the poor performance of space group prediction (see Table S6 in the Supplementary Material). Interestingly, the PEARSON model showed an appreciable increase to 83% when the top-3 accuracy (3 most probable classes) values are considered, while corresponding values for the COD and ICSD show smaller improvements.

3.5. Variable importance

The predictive power of the different variables was calculated from the decrease in accuracy associated with their exclusion. Bar plots in Figs. 12 (for crystal systems) and 13 (for space groups) show the calculated importance (scale of 0–100) of the 10 most influential variables in the models trained on the 5 datasets.

The Mendeleev number (`mendeleevnum`), which represents the similitude in chemical behaviour, played a key role in the classification of both crystal system and space group. Another set of descriptors based on the filled (`NdValence, NpValence`) and unfilled (`NdUnfValence, NpUnfValence`) *d* and *p* orbitals were found pivotal for multilabel classification. Variables such as work function, angular momentum quantum number (`lquant`) (which describes the orbital shape), first ionization energy, single-bond covalent radius [68] and heat of fusion represented less influential and generalized contributions. The present results show that variables, such as specific heat, atomic packing misfitting [69,70] and the electronegativity scale from Rahm et al. [45] which are not included in the Magpie [50] set of features are relevant for prediction of crystallographic information. In particular, the atomic packing misfitting was shown to be a critical variable for the HE compounds.

4. Conclusions

In this study we have assessed the predictive utility of machine learning models created from 5 different repositories of experimental or theoretical crystallographic data. For both crystal system and space group predictions, the performance shows strong dependence on the class distribution in the dataset used for training. In the validation tests conducted, the models trained on PEARSON (and to some extent on ICSD) were found to be more consistent and exhibiting better predictive ability across all other datasets and additional experimental data. This is attributed to a more balanced distribution of the classes compared to more skewed ratios in databases such as COD and OQMD.

Overall, the present work demonstrates that random forest models (in particular, the ones trained on the PEARSON dataset) for both multiclass and multilabel classification problems were able to capture decision rules that can facilitate rapid and directed

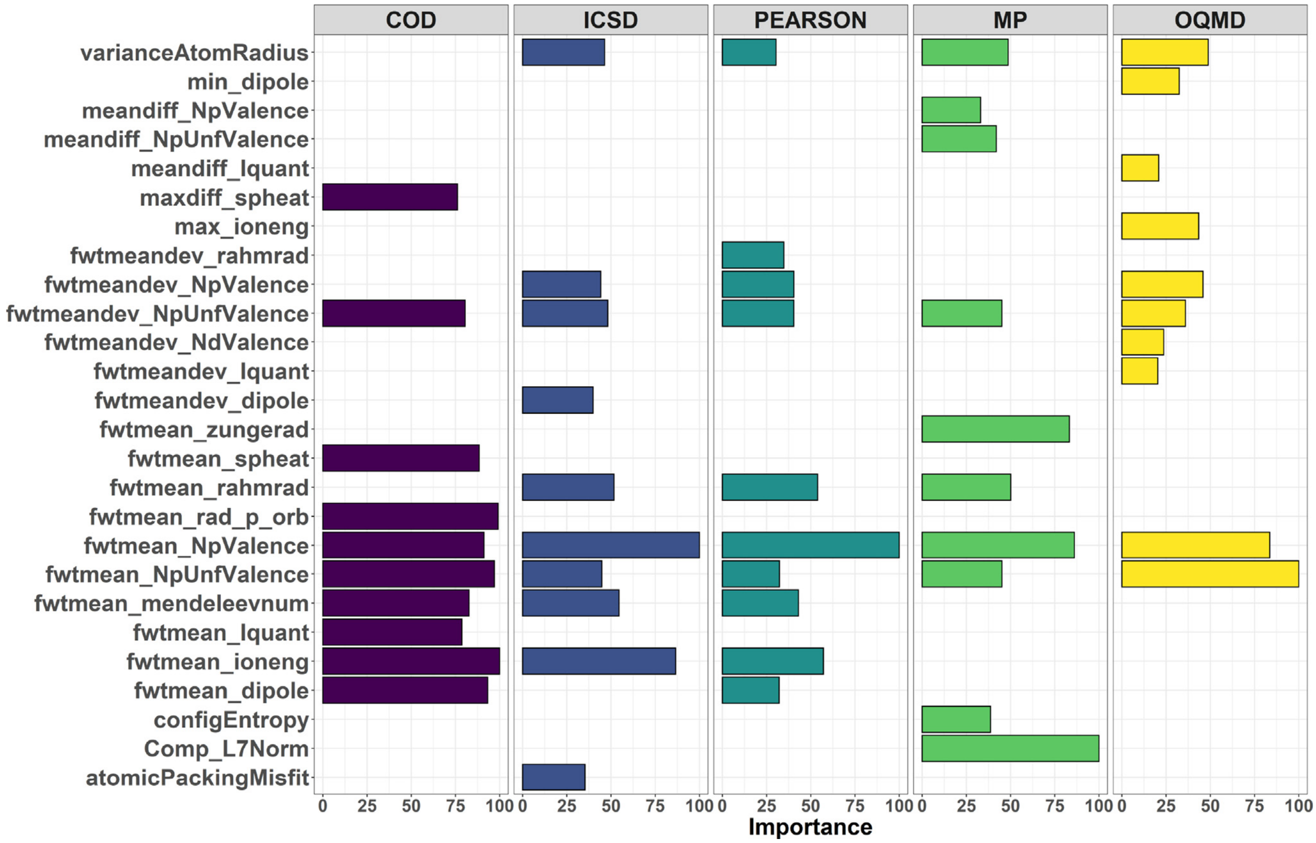


Fig. 12. Importance of the variables in the random forest models for crystal system prediction (importance normalized to 100). See Supplementary Material for the explanation of the variables.

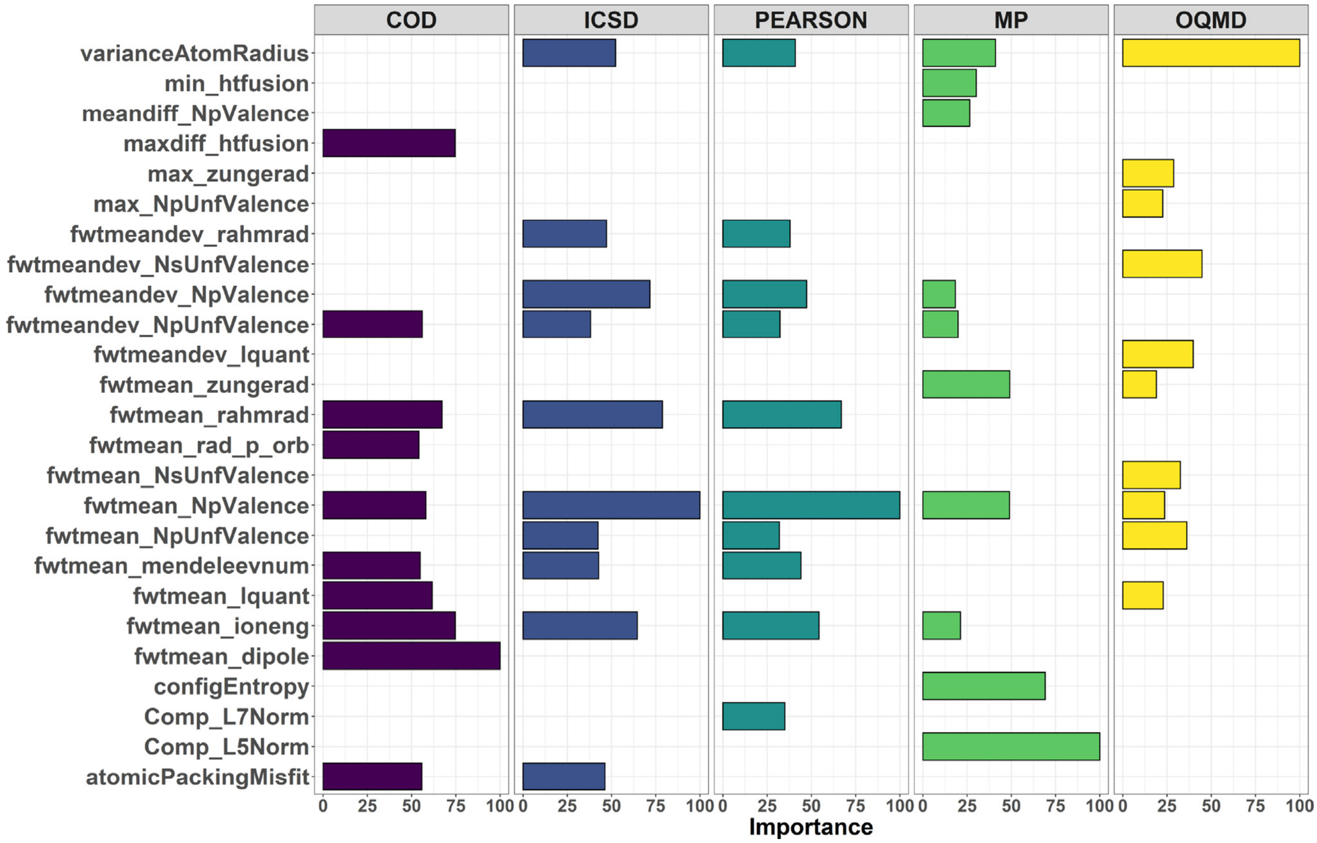


Fig. 13. Importance of the variables in the random forest models for space group prediction (importance normalized to 100). See Supplementary Material for the explanation of the variables.

design of new materials. Although the stoichiometry-based deep learning approach was shown to perform well on a number of regression tasks [31], its current application for multiclass classification has yielded only average results.

On a final note on the predictive ability of the existing databases, the fact that only about 50 space groups comprise sufficient data for successful machine learning shows that the curators of each theoretical and experimental database must develop specific efforts to populate the sparse classes, both to fulfill a sound information goal and because biased databases tend to delay the discovery of exotic materials with unusual space groups. Clearly, the materials science field and the world at large could greatly profit from a major merging operation of these partially incomplete and differently biased crystallographic databases.

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.

Acknowledgements

The authors gratefully acknowledge the support from the Research Council of Norway under Grant agreements Nos. 275752 and 289545.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.actamat.2022.118353

References

[1] A. Talapatra, S. Boluki, P. Honarmandi, A. Solomou, G. Zhao, S.F. Ghoreishi, A. Molkeri, D. Allaire, A. Srivastava, X. Qian, E.R. Dougherty, D.C. Lagoudas, R. Arróyave, Experiment design frameworks for accelerated discovery of targeted materials across scales, Front. Mater. 6 (2019), doi:10.3389/fmats.2019.00082.

[2] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (21) (2019) 1900808, doi:10.1002/advs.201900808.

[3] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: the materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, doi:10.1063/1.4812323.

[4] K. Choudhary, K.F. Garrity, A.C.E. Reid, B. DeCost, A.J. Biacchi, A.R.H. Walker, Z. Trautt, J. Hattrick-Simpers, A.G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S.V. Kalinin, B.G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe, F. Tavazza, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, Npj Comput. Mater. 6 (1) (2020), doi:10.1038/s41524-020-00440-1.

[5] S. Curtarolo, W. Setyawan, G.L. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, D. Morgan, AFLOW: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218–226, doi:10.1016/j.commatsci.2012.02.005.

[6] C. Draxl, M. Scheffler, The NOMAD laboratory: from data sharing to artificial intelligence, JPhys Mater. 2 (3) (2019) 036001, doi:10.1088/2515-7639/ab13bb.

[7] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, S. Rehme, Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features, J. Appl. Crystallogr. 52 (5) (2019) 918–925, doi:10.1107/s160057671900997x.

[8] A. Vaitkus, A. Merkys, S. Graulis, Validation of the crystallography open database using the crystallographic information framework, J. Appl. Crystallogr. 54 (2) (2021) 661–672, doi:10.1107/S1600576720016532.

[9] G. Hautier, Finding the needle in the haystack: materials discovery and design through computational ab initio high-throughput screening, Comput. Mater. Sci. 163 (2019) 108–116, doi:10.1016/j.commatsci.2019.02.040.

[10] N. Marzari, A. Ferretti, C. Wolverton, Electronic-structure methods for materials design, Nat. Mater. 20 (6) (2021) 736–749, doi:10.1038/s41563-021-01013-3.

[11] L. Sun, Y.-X. Zhou, X.-D. Wang, Y.-H. Chen, V.L. Deringer, R. Mazzarello, W. Zhang, Ab initio molecular dynamics and materials design for embedded phase-change memory, Npj Comput. Mater. 7 (1) (2021), doi:10.1038/s41524-021-00496-7.

[12] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, Npj Comput. Mater. 5 (1) (2019), doi:10.1038/s41524-019-0221-0.

[13] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, Machine learning assisted materials design and discovery for rechargeable batteries, Energy Storage Mater. 31 (2020) 434–450, doi:10.1016/j.ensm.2020.06.033.

[14] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, Machine learning-driven new material discovery, Nanoscale Adv. 2 (8) (2020) 3115–3130, doi:10.1039/d0na00388c.

[15] J.E. Saal, A.O. Oliynyk, B. Meredig, Machine learning in materials discovery: confirmed predictions and their underlying approaches, Annu. Rev. Mater. Res. 50 (1) (2020) 49–69, doi:10.1146/annurev-matsci-090319-010954.

[16] J. Graser, S.K. Kauwe, T.D. Sparks, Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons, Chem. Mater. 30 (11) (2018) 3601–3612, doi:10.1021/acs.chemmater.7b05304.

[17] A.R. Oganov, C.J. Pickard, Q. Zhu, R.J. Needs, Structure prediction drives materials discovery, Nat. Rev. Mater. 4 (5) (2019) 331–348, doi:10.1038/s41578-019-0101-8.

[18] D. Ma, B. Grabowski, F. Körmann, J. Neugebauer, D. Raabe, Ab initio thermodynamics of the CoCrFeMnNi high entropy alloy: importance of entropy contributions beyond the configurational one, Acta Mater. 100 (2015) 90–97, doi:10.1016/j.actamat.2015.08.050.

[19] K. Ryan, J. Lengyel, M. Shatruk, Crystal structure prediction via deep learning, J. Am. Chem. Soc. 140 (32) (2018) 10158–10168, doi:10.1021/jacs.8b03913.

[20] E.V. Podryabinkin, E.V. Tikhonov, A.V. Shapeev, A.R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, Phys. Rev. B 99 (6) (2019), doi:10.1103/physrevb.99.064114.

[21] H. Wang, Y. Zhang, L. Zhang, H. Wang, Crystal structure prediction of binary alloys via deep potential, Front. Chem. 8 (2020), doi:10.3389/fchem.2020.589795.

[22] C.J. Court, B. Yildirim, A. Jain, J.M. Cole, 3-D inorganic crystal structure generation and property prediction via representation learning, J Chem. Inf. Model. 60 (10) (2020) 4518–4535, doi:10.1021/acs.jcim.0c00464.

[23] Q. Tong, P. Gao, H. Liu, Y. Xie, J. Lv, Y. Wang, J. Zhao, Combining machine learning potential and structure prediction for accelerated materials design and discovery, J. Phys. Chem. Lett. 11 (20) (2020) 8710–8720, doi:10.1021/acs.jpclett.0c02357.

[24] T. Jin, I. Park, T. Park, J. Park, J.H. Shim, Accelerated crystal structure prediction of multi-elements random alloy using expandable features, Sci. Rep. 11 (1) (2021), doi:10.1038/s41598-021-84544-8.

[25] C.-H. Liu, Y. Tao, D. Hsu, Q. Du, S.J.L. Billinge, Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function, Acta Crystallogr. A 75 (4) (2019) 633–643, doi:10.1107/s2053273319005606.

[26] H. Liang, V. Stanev, A.G. Kusne, I. Takeuchi, CRYSPNet: crystal structure predictions via neural networks, Phys. Rev. Mater. 4 (12) (2020), doi:10.1103/physrevmaterials.4.123802.

[27] Y. Zhao, Y. Cui, Z. Xiong, J. Jin, Z. Liu, R. Dong, J. Hu, Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions, ACS Omega 5 (7) (2020) 3596–3606, doi:10.1021/acsomega.9b04012.

[28] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation, Mach. Learn. Sci. Technol. 1 (4) (2020) 045024, doi:10.1088/2632-2153/aba947.

[29] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572, doi:10.1021/acs.chemmater.9b01294.

[30] A.N. Zaloga, V.V. Stanovov, O.E. Bezrukova, P.S. Dubinin, I.S. Yakimov, Crystal symmetry classification from powder x-ray diffraction patterns using a convolutional neural network, Mater. Today Commun. 25 (2020) 101662, doi:10.1016/j.mtcomm.2020.101662.

[31] R.E.A. Goodall, A.A. Lee, Predicting materials properties without crystal structure: deep representation learning from stoichiometry, Nat. Commun. 11 (1) (2020), doi:10.1038/s41467-020-19964-7.

[32] S. Kong, D. Guevarra, C.P. Gomes, J.M. Gregoire, Materials representation and transfer learning for multi-property prediction, Appl. Phys. Rev. 8 (2) (2021) 021409, doi:10.1063/5.0047066.

[33] Y. Li, R. Dong, W. Yang, J. Hu, Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors, Comput. Mater. Sci. 198 (2021) 110686, doi:10.1016/j.commatsci.2021.110686.

[34] Y. Li, W. Yang, R. Dong, J. Hu, Mlatticeabc: generic lattice constant prediction of crystal materials using machine learning, ACS Omega 6 (17) (2021) 11585–11594, doi:10.1021/acsomega.1c00781.

[35] A.Y.-T. Wang, S.K. Kauwe, R.J. Murdock, T.D. Sparks, Compositionally restricted attention-based network for materials property predictions, Npj Comput. Mater. 7 (1) (2021), doi:10.1038/s41524-021-00545-1.

[36] Pearson's crystal data: crystal structure database for inorganic compounds (on dvd), release 2020/21, 2021, (ASM International, Materials Park, Ohio, USA). [Accessed: August-2021].

[37] 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 (11) (2013) 1501–1509, doi:10.1007/s11837-013-0755-4.

[38] J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau, S.-Y. Chang, Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes, Adv. Eng. Mater. 6 (5) (2004) 299–303, doi:10.1002/adem.200300567.

[39] B. Cantor, I. Chang, P. Knight, A. Vincent, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng. A 375–377 (2004) 213–218, doi:10.1016/j.msea.2003.10.257.

[40] E.P. George, D. Raabe, R.O. Ritchie, High-entropy alloys, Nat. Rev. Mater. 4 (8) (2019) 515–534, doi:10.1038/s41578-019-0121-4.

[41] C. Oses, C. Toher, S. Curtarolo, High-entropy ceramics, Nat. Rev. Mater. 5 (4) (2020) 295–309, doi:10.1038/s41578-019-0170-8.

[42] M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, A. Vaitkus, Using SMILES strings for the description of chemical connectivity in the crystallography open database, J. Cheminf. 10 (1) (2018), doi:10.1186/s13321-018-0279-6.

[43] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N.R. Serebryanaya, P. Moeck, R.T. Downs, A.L. Bail, Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration, Nucl. Acids Res. 40 (D1) (2011) D420–D427, doi:10.1093/nar/gkr900.

[44] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies, Npj Comput. Mater. 1 (1) (2015), doi:10.1038/npjcompumats.2015.10.

[45] M. Rahm, T. Zeng, R. Hoffmann, Electronegativity seen as the ground-state average valence electron binding energy, J. Am. Chem. Soc. 141 (1) (2018) 342–351, doi:10.1021/jacs.8b10246.

[46] A. Zunger, Systematization of the stable crystal structure of allAB-type binary compounds: a pseudopotential orbital-radii approach, Phys. Rev. B 22 (12) (1980) 5839–5872, doi:10.1103/physrevb.22.5839.

[47] L. Mentel, mendeleev – a python resource for properties of chemical elements, ions and isotopes, 2021. https://github.com/lmmentel/mendeleev.

[48] V. Venkatraman, The utility of composition-based machine learning models for band gap prediction, Comput. Mater. Sci. 197 (2021) 110637, doi:10.1016/j.commatsci.2021.110637.

[49] S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, F. Pan, Encoding the atomic structure for machine learning in materials science, WIRES Comput. Mol. Sci. (2021), doi:10.1002/wcms.1558.

[50] 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), doi:10.1038/npjcompumats.2016.28.

[51] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, doi:10.1023/a:1010933404324.

[52] M.N. Wright, A. Ziegler, Ranger: a fast implementation of random forests for high dimensional data in C++ and R, J. Stat. Soft. 77 (1) (2017) 1–17, doi:10.18637/jss.v077.i01.

[53] R Core Team, R: A language and environment for statistical computing, 2020, (Version 4.2.1). https://www.R-project.org/.

[54] V. Revi, S. Kasodariya, A. Talapatra, G. Pilania, A. Alankar, Machine learning elastic constants of multi-component alloys, Comput. Mater. Sci 198 (2021) 110671, doi:10.1016/j.commatsci.2021.110671.

[55] V. Venkatraman, FP-ADMET: a compendium of fingerprint-based ADMET prediction models, J. Cheminf. 13 (1) (2021), doi:10.1186/s13321-021-00557-5.

[56] T. Zhang, J. Su, Z. Xu, Y. Luo, J. Li, Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier, Appl. Sci. 11 (2) (2021) 543, doi:10.3390/app11020543.

[57] J. Quist, L. Taylor, J. Staaf, A. Grigoriadis, Random forest modelling of high-dimensional mixed-type data for breast cancer classification, Cancers 13 (5) (2021) 991, doi:10.3390/cancers13050991.

[58] H. Ishwaran, T.A. Gerds, U.B. Kogalur, R.D. Moore, S.J. Gange, B.M. Lau, Random survival forests for competing risks, Biostatistics 15 (4) (2014) 757–773, doi:10.1093/biostatistics/kxu010.

[59] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, Z.M. Jones, mlr: Machine learning in R, J Mach. Learn. Res. 17 (170) (2016) 1–5.

[60] P. Probst, Q. Au, G. Casalicchio, C. Stachl, B. Bischl, Multilabel classification with R package mlr, R J. 9 (1) (2017) 352–369, doi:10.32614/RJ-2017-012.

[61] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process Manag. 45 (4) (2009) 427–437, doi:10.1016/j.ipm.2009.03.002.

[62] C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognit. Lett. 30 (1) (2009) 27–38, doi:10.1016/j.patrec.2008.08.010.

[63] E. Rendón, R. Alejo, C. Castorena, F.J. Isidro-Ortega, E.E. Granda-Gutiérrez, Data sampling methods to deal with the big data multi-class imbalance problem, Appl. Sci. 10 (4) (2020) 1276, doi:10.3390/app10041276.

[64] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, doi:10.1145/2939672.2939785.

[65] Y. Li, R. Dong, W. Yang, J. Hu, Composition based crystal materials symmetry prediction using machine learning with enhanced descriptors, Comput. Mater. Sci. 198 (2021) 110686, doi:10.1016/j.commatsci.2021.110686.

[66] R. Jaafreh, T. Abuhmed, J.-G. Kim, K. Hamad, Crystal structure guided machine learning for the discovery and design of intrinsically hard materials, J. Mater. 8 (3) (2022) 678–684, doi:10.1016/j.jmat.2021.11.004.

[67] A.A. Alsaui, Y.A. Alghofaili, M. Alghadeer, F.H. Alharbi, Resampling techniques for materials informatics: limitations in crystal point groups classification, J. Chem. Inf. Model. 62 (15) (2022) 3514–3523, doi:10.1021/acs.jcim.2c00666.

[68] P. Pyykkö, M. Atsumi, Molecular single-bond covalent radii for elements 1–118, Chem. Eur. J. 15 (1) (2009) 186–197, doi:10.1002/chem.200800987.

[69] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J Appl. Phys. 109 (10) (2011) 103505, doi:10.1063/1.3587228.

[70] Z. Wang, Y. Huang, Y. Yang, J. Wang, C. Liu, Atomic-size effect and solid solubility of multicomponent alloys, Scr. Mater. 94 (2015) 28–31, doi:10.1016/j.scriptamat.2014.09.010.