JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

Accurate space-group prediction from composition

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

[a]Norwegian University of Science and Technology, 7491 Trondheim, Norway, [b]SINTEF Materials Physics, 0373 Oslo, Norway, and [c]CEFEMA, Instituto Superior Tecnico, University of Lisbon, Lisbon, Portugal. *Correspondence e-mail: vishwesh.venkatraman@ntnu.no, patricia.carvalho@sintef.no

Received 19 November 2023
Accepted 14 May 2024

Edited by P. Munshi, Shiv Nadar Institution of Eminence, Gautam Buddha Nagar, India

Keywords: space groups; random forests; machine learning; data sets; prediction.

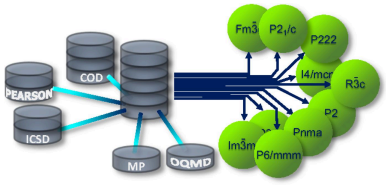
Supporting information: this article has supporting information at journals.iucr.org/j

Predicting crystal symmetry simply from chemical composition has remained challenging. Several machine-learning approaches can be employed, but the predictive value of popular crystallographic databases is relatively modest due to the paucity of data and uneven distribution across the 230 space groups. In this work, virtually all crystallographic information available to science has been compiled and used to train and test multiple machine-learning models. Composition-driven random-forest classification relying on a large set of descriptors showed the best performance. The predictive models for crystal system, Bravais lattice, point group and space group of inorganic compounds are made publicly available as easy-to-use software downloadable from https://gitlab.com/vishsoft/cosy.

1. Introduction

Exhaustive screening via experimentation is prohibitive and identification of interesting compounds is increasingly reliant on efficient computational methods, such as high-throughput *ab initio* simulations (Hautier, 2019; Marzari *et al.*, 2021; Sun *et al.*, 2021) and machine learning (ML) (Axelrod *et al.*, 2022; Kusaba *et al.*, 2022; Venkatraman & Carvalho, 2022). Yet, establishing crystallographic information from first principles remains computationally expensive and requires an educated guess on candidate structures (Oganov *et al.*, 2019). On the other hand, ML methods demand large amounts of data and, although popular crystallographic databases have been used before, the predictions do not meet the expectations for practical use (Venkatraman & Carvalho, 2022).

Various strategies for structural representation have been adopted within ML, ranging from graph-based methodologies (Chen *et al.*, 2019) to atomic pair distribution functions (PDFs) (Liu *et al.*, 2019). In the latter approach, a convolutional neural network model achieved top-6 prediction accuracy of $\simeq$90% for 45 heavily populated space groups (Liu *et al.*, 2019). Others have relied directly on powder X-ray diffraction patterns combined with data augmentation through simulation (Park *et al.*, 2017; Oviedo *et al.*, 2019; Suzuki *et al.*, 2020). In this context, the ensemble decision-tree model proposed by Suzuki *et al.* (2020) attained accuracies exceeding 90% for crystal system classification and surpassed 80% top-5 accuracy for space-group prediction. Structure-agnostic methodologies attempting to learn patterns/associations purely from compound composition have also been employed, often through deep learning (Liang *et al.*, 2020; Goodall & Lee, 2020; Kong *et al.*, 2021; Li *et al.*, 2021*a*,*b*). Against this background, Venkatraman & Carvalho (2022) have shown that, for sufficiently large and well represented databases, random-forest (RF) models using composition-based descriptors trained on the 50



most frequent space groups consistently result in better predictions than those obtained by deep learning, with 0.75–0.34 versus 0.65–0.25 F1 scores, respectively, depending on the particular crystallographic database employed to train the models.

All previous ML studies have used very limited data sets or focused on only around 50 heavily populated space groups due to paucity of data and class skewness of the existing crystallographic databases (Venkatraman & Carvalho, 2022; Liu *et al.*, 2019). In this work, we have augmented the data and enhanced class coverage by compiling virtually all crystallographic information available to science on inorganic compounds. Composition-driven models based on demonstrated ML approaches (Saal *et al.*, 2020; Alsaui *et al.*, 2022; Venkatraman & Carvalho, 2022) have been trained on this merged database (MERGED) and tested for predicting crystal system, cell centering, Bravais lattice, point group and space group. The models have been integrated into user-friendly public domain software to facilitate fast prediction of crystal symmetry.

2. Data sets

Data were compiled from three experimental databases, namely, the Crystallography Open Database (COD) (Vaitkus *et al.*, 2021), Pearson's Crystal Data (PEARSON) (ASM International, 2021) and the Inorganic Crystal Structure Database (ICSD) (Zagorac *et al.*, 2019), and combined with data from two databases containing structures calculated with density functional theory (DFT), namely, Open Quantum Materials Database (OQMD) (Saal *et al.*, 2013) and Materials Project (MP) (Jain *et al.*, 2013).

For each of the primary data sets, the numbers in the chemical formulas have been normalized to 1 and rounded to the fourth decimal position. Compounds for which the formulas could not be parsed, as well as duplicate entries, were eliminated. Structures comprising only a single element or noble gas(es) were also excluded.

Polymorphism is exhibited by fewer than 8% of the compounds in experimental databases. Therefore, meaningful predictions for these multi-labeled entries are hindered by data scarcity, and thus compounds crystallizing in multiple space groups have been excluded from the experimental data sets. A comprehensive treatment of multi-labeled data in the context of polymorphism can be found elsewhere (Venkatraman & Carvalho, 2022).

Given the high proportion of poorly represented space groups across all repositories, experimental and theoretical databases were merged to augment the data. While stability conditions for compounds in experimental repositories are often not readily available, it can generally be assumed that compounds adopting a single structure have been solved under standard atmospheric conditions of temperature and pressure, rendering them stable or prevalent at such conditions. For theoretical databases, only entries with a stability indicator of $E_{\rm hull} = 0$ were retained. The decision to exclude

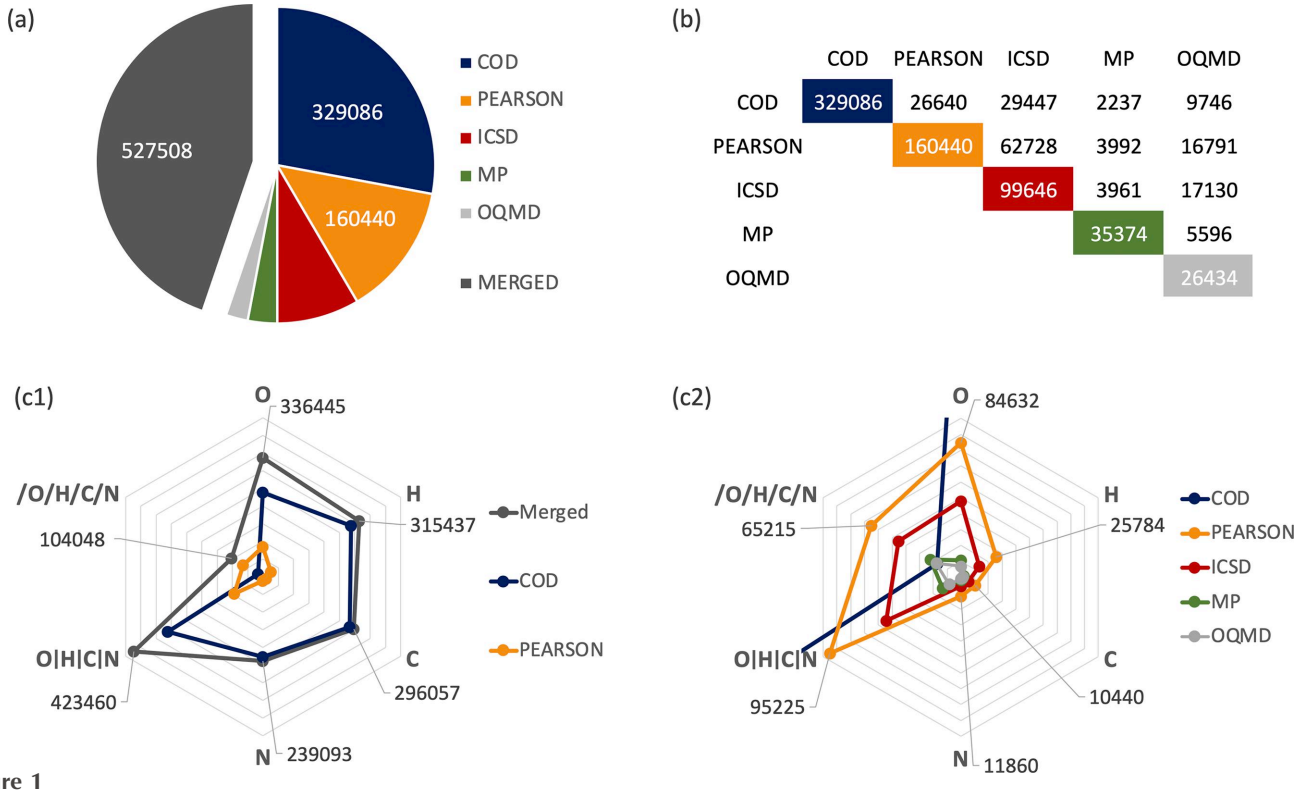


Figure 1
(*a*) Distribution of unique compounds in the databases. (*b*) Pairwise intersections. (*c*1) Element-containing compounds in each database, where O|H|C|N represents compounds containing at least one of these elements and /O/H/C/N represents compounds containing none of these elements. (*c*2) Magnified detail of (*c*1).

compounds with multiple stability domains from experimental repositories and those lying above the DFT convex hull from theoretical repositories enhanced the proportion of compounds stable at moderate temperatures. This approach established a common foundation for both types of data.

Fig. 1(*a*) represents the number of unique compounds in each database. After duplicate removal, the merged database comprised 527 508 unique compounds spanning 87 elements [see Table S1 in the supporting information (SI)]. Most data originate from the experimental databases, with the theoretical repositories offering less than 12% of the total number of compounds. Fig. 1(*b*) lists the pairwise intersections between the different data sets. Significant overlap exists between the experimental commercial data sets (PEARSON contains 63% of ICSD and ICSD contains 39% of PEARSON) and between these and OQMD (ICSD contains 65% of OQMD and PEARSON contains 64% of OQMD).

Actinium and polonium are the least common of the 87 elements present in MERGED (respectively, 0.06 and 0.006% of compounds). On the other hand, most compounds contain light elements such as oxygen (64%), hydrogen (60%), carbon (56%) and nitrogen (45%). Their distribution across the different databases is shown in Figs. 1(*c*1) and 1(*c*2). Compounds with at least one of the frequent light elements, *i.e.* O|H|C|N, account for 95% of the entries in COD, 59% in PEARSON, 55% in ICSD, 37% in MP and 32% in OQMD which, after duplicate removal, yielded 80% in MERGED. This proportion, significantly lower than that in COD, reflects the contribution to diversity imparted by the smaller databases with higher fractions of compounds without the frequent light elements [/O/H/C/N, see Figs. 1(*c*1) and 1(*c*2)]. Indeed, the large COD repository comprises a high fraction of mineral structures, which typically contain light elements, whereas the other primary databases seem more application oriented.

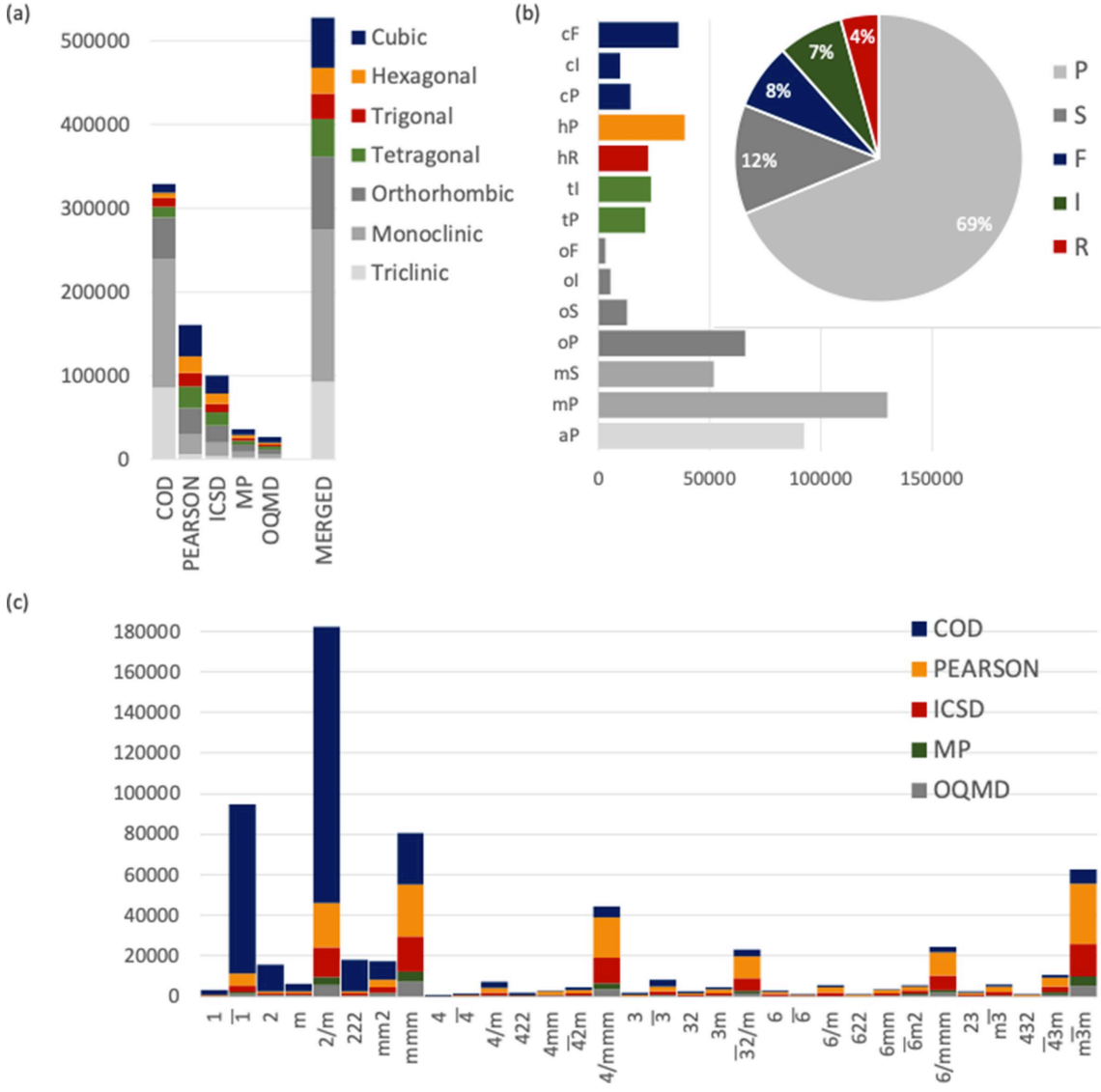


Figure 2
(*a*) Compound distribution across the crystal systems in all databases. (*b*) Distribution of compounds across the Bravais lattices and lattice centering types in MERGED. (*c*) Contribution of each primary database to each point-group class.

Fig. 2 breaks down the compound distribution across (*a*) the seven crystal systems for all databases, (*b*) the 14 Bravais lattices and five lattice centering types for MERGED, and (*c*) the 32 point groups for the primary databases. The monoclinic crystal system is the dominant class in MERGED, followed by the orthorhombic and triclinic systems. The predominance of these systems is essentially inherited from COD [see Fig. 2(*a*)]. The other primary databases show more balanced crystal system distributions and therefore contribute to enhancing the tetragonal, trigonal, hexagonal and cubic classes in MERGED [see Fig. 2(*a*)]. The dominant centering type in MERGED is *P* due to the high number of compounds with triclinic, monoclinic and orthorhombic primitive lattices, while the remaining centering types show more balanced proportions [see Fig. 2(*b*)]. In the primary databases each crystal system exhibits a clearly preponderant point group: $\overline{1}$ for triclinic, 2/*m* for monoclinic, *mmm* for orthorhombic, $\overline{4}/mmm$ for tetragonal, $\overline{3}2/m$ for trigonal, 6/*mmm* for hexagonal and $m\overline{3}m$ for cubic [see Fig. 2(*c*); a similar overall point-group distribution was obtained for MERGED (as shown in Fig. F1 of the SI), *i.e.* duplicate removal did not change the preponderant point groups in each crystal system].

The heatmap in Fig. 3 reveals important differences between the primary databases in terms of space-group

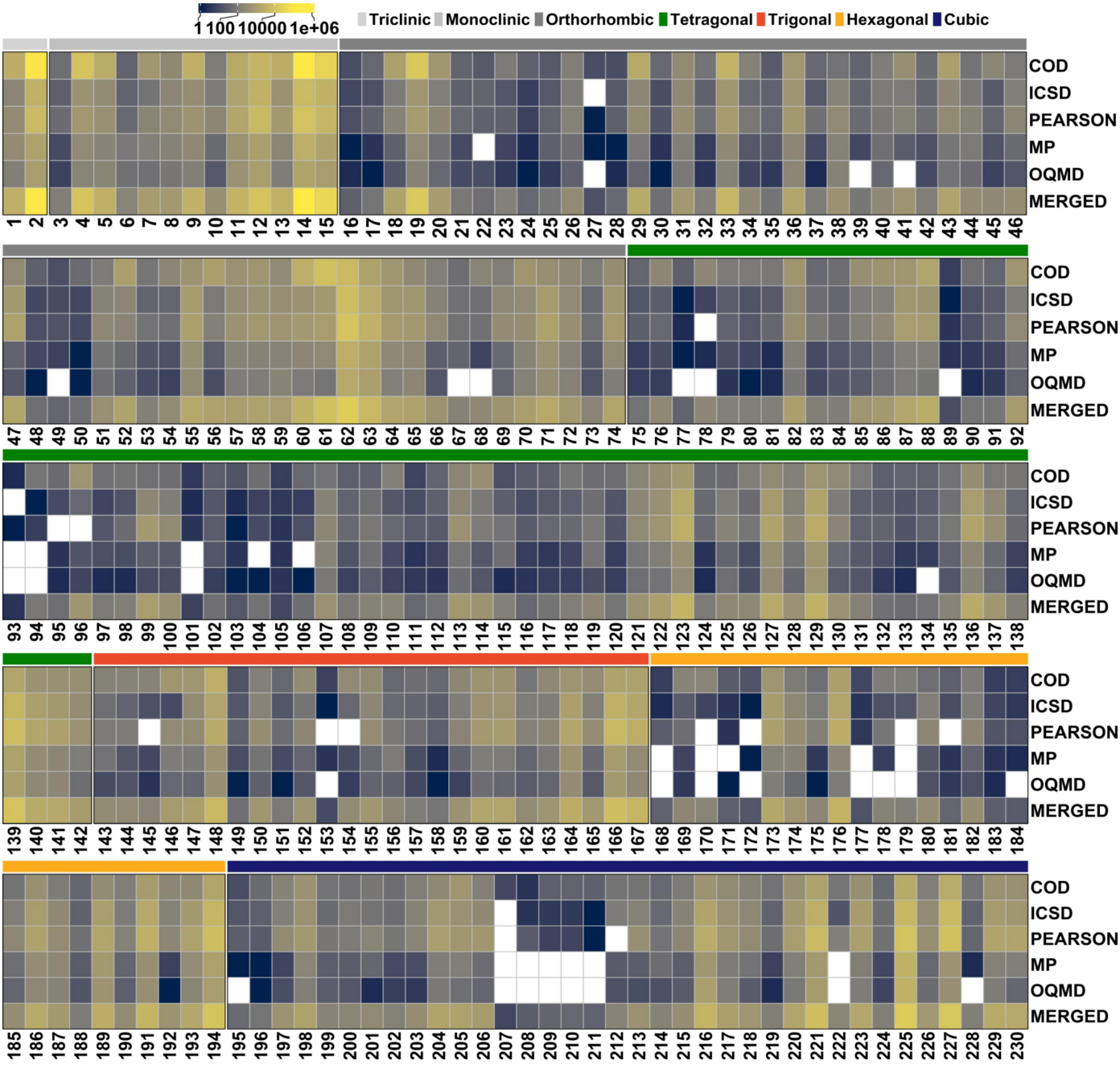


Figure 3
Distribution of unique compounds across the 230 space groups in the data sets.

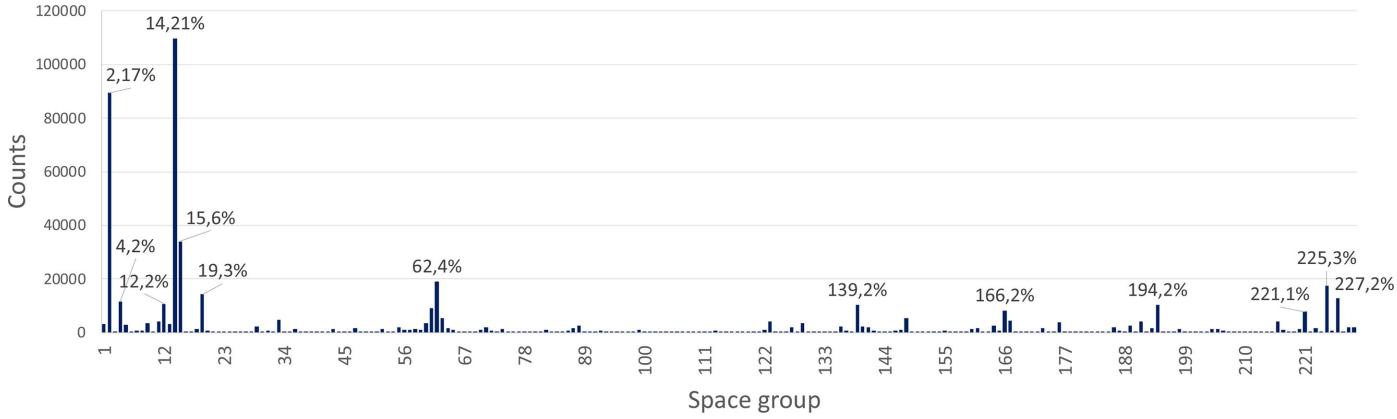


Figure 4
Space-group distribution in MERGED. The labels over frequent classes indicate the space-group number followed by the corresponding percentage. The number of compounds for each space group is listed in Table S2 in the SI.

representation. MERGED shows an overall improved coverage, with nearly 75% of the space groups comprising more than 100 compounds. Nonetheless, five space groups, namely, 93, 101, 105, 153 and 207, still contain fewer than ten compounds in the merged database. Meanwhile, space group 14 is the most populated, with over 100 000 compounds originating mostly from COD, but with significant contributions from PEARSON and ICSD. Despite the variations, all primary databases show highly frequent 2, 14, 15, 62, 225 and 227 space groups (see Fig. 3). The histogram in Fig. 4 further illustrates the uneven class occupation in MERGED. The striking preference of crystallization for specific space groups, leading to extremely frequent versus rare space groups, is a well known fact but a complex problem in crystallography and, to the best of the authors' knowledge, clear space-group filling rules are yet to be defined for inorganic compounds [see the review by Urusov & Nadezhina (2009)].

3. Modeling

In a previous study (Venkatraman & Carvalho, 2022), the predictive performance of an ensemble decision-tree RF approach was compared against a deep-learning framework named Roost (Representation Learning from Stoichiometry), reported by Goodall & Lee (2020) to outperform even ElemNet, an alternative neural network method (Jha *et al.*, 2018). For each of the primary databases, the RF approach showed an overall better performance in multi-class classification (Venkatraman & Carvalho, 2022). Notably, Roost yielded disappointing results except for OQMD, which is severely skewed towards cubic structures, fostering cubic predictions and thereby justifying the earlier reports [only OQMD was employed by Goodall & Lee (2020) to test Roost and by Jha *et al.* (2018) to test ElemNet]. Here, for the augmented data set MERGED, we expand the palette of tools and analyze the prediction performance of

(i) the ensemble decision-tree RF approach relying on a large set of composition-based descriptors, which has been effectively tested in previous studies (Venkatraman, 2021, 2023; Venkatraman & Carvalho, 2022),

against two other ML approaches:

(ii) DUET, which combines bagging with boosting decision-tree methods via two classifiers (Vargaftik & Ben-Itzhak, 2022). A bagging model (RF-based) is trained on the entire training data set and, subsequently, a boosting model (XGBoost) is trained on a fraction of the data set for which the bagging model underperformed. To rank how valuable a given labeled sample is to training the boosting classifier, a heuristic called 'data instance predictability' is used to define the data set fraction for training the boosting model. The predictability-driven fraction of the training data set was set to the recommended value of 6% (Vargaftik & Ben-Itzhak, 2022).

(iii) TabNet, a deep-learning architecture for tabular data (such as a matrix of descriptor vectors) that uses sequential attention to select features from which to reason at each decision step (Arik & Pfister, 2019).

For each of the three approaches employed, the compounds were split into calibration (80%) and test (20%) sets. For each compound, a descriptor vector based on maximum, minimum, fraction-weighted mean and mode, as well as average deviations of the elemental properties (such as electronegativity, atomic weight, polarizability and number of filled/unfilled valence orbitals), was calculated using software written in Java (available from https://github.com/vvishwesh/MaterialDescriptors). The descriptor set includes other variables derived from element properties, such as specific heat and atomic packing efficiency (Guo *et al.*, 2011), as well as different electronegativity scales (Rahm *et al.*, 2019), not included in the original Magpie set (Ward *et al.*, 2016). The descriptor vector contained missing values due to the non-availability of complete sets of elemental attributes. Thus, a cleaning step was applied where descriptors with missing values were removed. This was followed by a correlation-based variable reduction step to exclude highly correlated variables (a pairwise squared correlation cutoff of 0.90 was used), which resulted in data sets with 126–129 descriptors (the variation stems from the random selection of the train/test sets). Finally, a fivefold cross-validation was carried out to assess the generalizability of the models (potential performance on

unseen data). This was repeated three times with independent train/test splits to examine performance variability.

Space-group classes with less than 100 compounds have been excluded from the evaluation. The number of classes was hence seven for crystal systems, five for lattice centering, 14 for Bravais lattices, 32 for point groups and 172 for space groups (since in MERGED still 58 of the 230 space groups comprised less than 100 compounds). The assessment of the models was based on top-k accuracy, *i.e.* on the proportion of compounds for which a correct answer is present in the top-k results. Here, we have evaluated the accuracy for k = 1, 2, 3, 5.

4. Results and discussion

4.1. Cross-validation

Fig. 5 provides a visual summary of the models' performance in predicting lattice centering, crystal system, Bravais lattice, point group and space group for the test sets. As expected, the overall prediction quality decreases with the number of classes (in brackets). The RF model achieved the highest top-k accuracies, followed by DUET. The accuracies for TabNet were considerably lower, except for crystal system prediction where the values are marginally close to those of the other approaches. The performance gaps are more significant for space-group prediction, with top-5 accuracy of only 0.43 $\pm$ 0.03 for TabNet. In contrast, RF achieved top-3 accuracy of 0.81 $\pm$ 0.001 and above in all symmetry classifications (see labels in Fig. 5; the standard deviations for each response can be found in Table S4 in the SI). Baseline performance (wherein no predictors were used and instead the target values were averaged in some way) was assessed using the *basemodels* package in R through dummy classifiers on the following basis: (i) the most frequent class in the training set was selected for all instances, and (ii) class labels were assigned according to the class distribution in the training set. The multi-class Cohen's kappa (Artstein & Poesio, 2008) values for these two dummy models across all training data sets were found to be close to 0. In comparison, the RF models exhibited values in the 0.5 to 0.6 range, confirming their robustness and generalization capability.

In order to understand the reasons behind the relatively low top-1 accuracy for space-group prediction, we examined the class-wise performance of the RF model. The per-class sensitivity and specificity variations (Fig. 6) show that the model typically has high specificity but low sensitivity. This variability in performance is seen, in particular, for the monoclinic (space-group numbers 3–9), orthorhombic (18–23, 29, 52, 56, 60, 61) and tetragonal (76–82, 96, 118, 119) crystal systems. The poor discrimination power can be attributed to the paucity of data for some space groups (less than 200 compounds for space groups 3, 6 and 32, for instance). Class imbalance is also an issue for accuracy which, as an evaluation metric, is more meaningful when the class labels are uniformly distributed.

Overall, both the TabNet and DUET models yielded poorer predictive performance than the RF-based models. For the

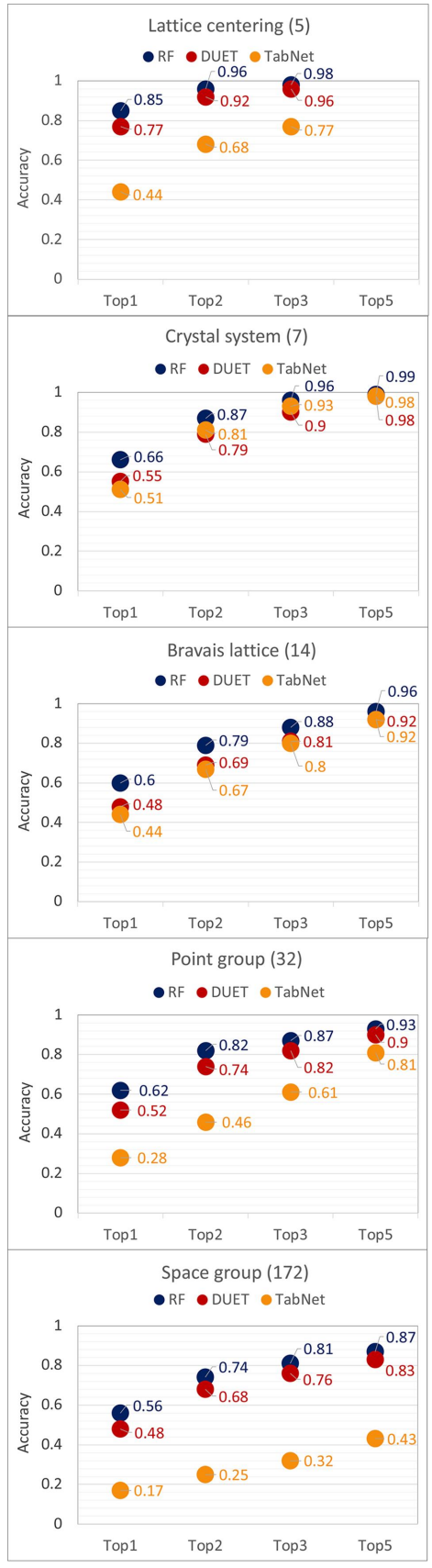


Figure 5
Top-k accuracies for the test set (averaged over three independent splits) obtained by the different ML approaches. Values in brackets indicate the number of classes associated with each response.

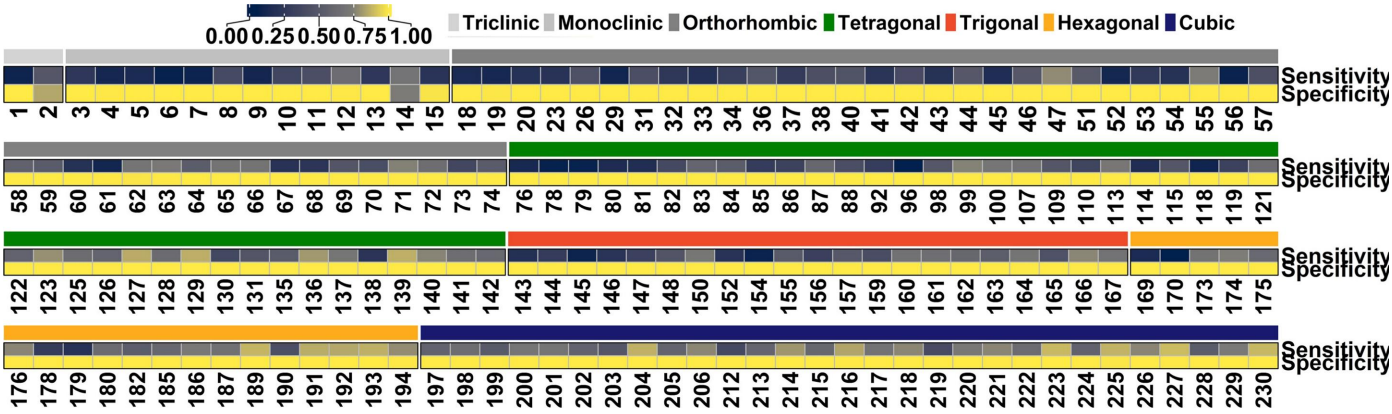


Figure 6
Heatmap showing the per-class sensitivity and specificity of the RF model for the 172 space-group test set.

TabNet models in particular, studies by Kadra *et al.* (2021), Shwartz-Ziv & Armon (2022) and Grinsztajn *et al.* (2022) have shown that methods such as TabNet or other deep tabular data modeling approaches do not outperform popular ensemble approaches such as XGBoost and other tree-based models. Importantly, Kadra *et al.* report that, for almost 40 different types of tabular data sets, decision-tree models were seen to perform strongly even against specialized neural architectures.

4.2. External validation

While the RF models exhibit high accuracy when predicting the symmetry of test sets, the performance for completely unseen data may provide more realistic estimates. To this end, we compiled two independent data sets: (i) compounds extracted from the *American Mineralogist* crystal structure database (AMCSD) (Downs & Hall-Wallace, 2003) and (ii) data on high-entropy alloys and compounds (HEAC) collated from multiple sources (see Table S3 in the SI). The validation data sets were prepared as described for the primary databases.

AMCSD comprises 8253 compounds not present in MERGED that could be employed for external validation. However, since 85 of them belong to space groups outside the 172 classes on which the model was trained, only 8168 compounds (distributed across 160 space groups) have been used for validation of space-group prediction. Nevertheless, the entire disjoint set has been considered for the other symmetry categories. The HEAC set is limited to 125 novel compounds restricted to 15 space groups (albeit highly skewed towards 194 and 225, see Table S3 in the SI), all of which are included in the 172 classes used to train the space-group model.

The top-*k* accuracies achieved in predicting the symmetry of the external data sets are presented in Fig. 7. For lattice centering, crystal system, Bravais lattice and point group, the predictions are highly accurate. Yet, the top-1 and top-2 performances are typically better for AMCSD than for HEAC. The space-group model yielded excellent metrics for AMCSD, with top-*k* accuracies exceeding 0.9 for all outcomes, while consistently lower performance was obtained for HEAC. Rather than only *per se* or against each other, the results achieved with external validations should also be evaluated in terms of the performance attained with the test sets (compare Fig. 7 with Fig. 5). In this scenario, several aspects are worth consideration:

(i) In the HEAC data set, each entry consists of five or more elements. The results achieved suggest that this characteristic is sufficiently well represented in the models. Indeed, MERGED comprises a relatively high fraction of compounds with more than four elements (55%, see Fig. F2 in the SI).

(ii) The AMCSD and HEAC data sets are comprised of, respectively, 10 and 83% of compounds without the prevalent light elements, while MERGED contains 20% of the /O/H/C/N compounds (see Fig. 1). The external validation shows that the amount of training data allowed the models to capture the crystallization behavior of both extremes, although this factor is likely to have contributed to the poorer prediction of the HEAC symmetries.

(iii) The external validation was carried out by training the models on all the data in MERGED, while only 80% was employed for internal validation. However, as demonstrated by the small standard deviations obtained for the three test splits (see Table S4 in the SI), the improvements with the training set augmentation are expected to be residual.

(iv) The AMCSD data set comprises more cubic and hexagonal and fewer triclinic compounds than MERGED, while the other crystal systems show similar distributions (see Fig. F5 in the SI). The HEAC data set consists essentially of high-symmetry compounds (89% cubic and hexagonal, see Table S3 in the SI). The better performance achieved with external data than with the test sets, particularly for HEAC, suggests that the models may better recognize the descriptor patterns associated with high symmetry.

(v) The very high accuracies obtained for AMCSD, particularly for space-group classification, indicate that the external compounds may be stoichiometrically similar to those present in MERGED. The effects of data augmentation on AMCSD prediction can clearly be appreciated by comparing the confusion matrices (top-1 accuracy) when training the

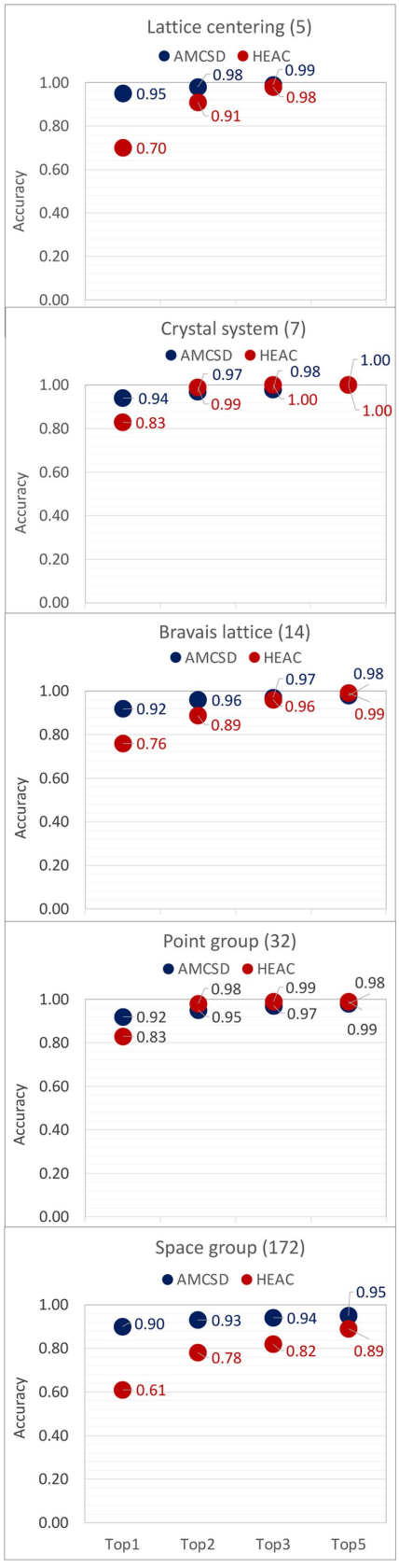


Figure 7
Top-*k* accuracies of the RF-based models for two independent test sets: (i) the AMCSD (Downs & Hall-Wallace, 2003) data set containing 8253 compounds and (ii) the HEAC data set comprising 125 compounds.

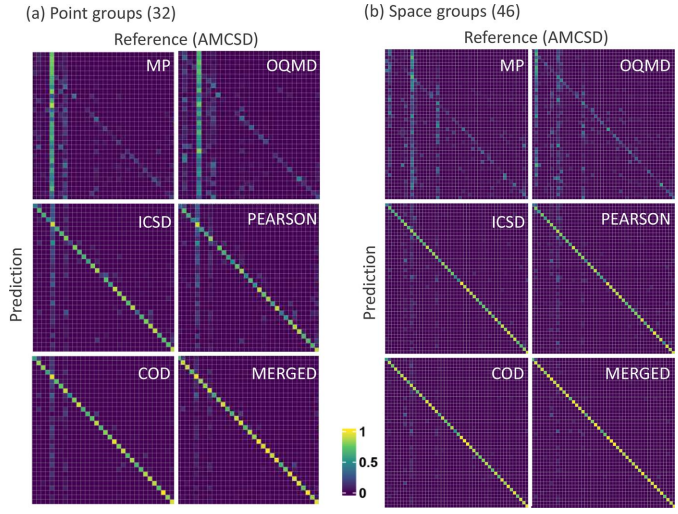


Figure 8
Confusion matrices for symmetry prediction using RF-based models trained on all databases (top-1 accuracy). (*a*) Point groups of the compounds in AMCSD. (*b*) Space groups of the compounds in AMCSD (for visualization clarity the RF-based models were trained only on the top 46 space groups of each database). Additional details can be found in Figs. F3 and F4 in the SI.

models on the different primary databases and on MERGED (see Fig. 8). The performance of models trained on COD and MERGED is high and similar as expected for mineral data. The stoichiometry similarity was further tested by changing the decimal position when rounding the numbers in the chemical formulas during the data preparation step. Rounding to the third decimal position reduced the AMCSD compounds not present in MERGED to 4374, while rounding to the second decimal position reduced the number of dissimilar compounds to 1320. The decimal position of rounding is however not a trivial matter since, for critical elements, precision is required to define the crystal structure adopted. Clearly, much work remains to be done in crystallographic data curation.

In summary, this work shows that high accuracy in symmetry prediction can be achieved by a decision-tree-based approach using solely elemental composition and the crystallographic information already available to science. In fact, the quality of the composition-driven prediction is notably higher than that obtained with models based on X-ray diffraction data (Suzuki *et al.*, 2020; Aguiar *et al.*, 2020; Corriero *et al.*, 2023) and atomic PDFs (Liu *et al.*, 2019), and also higher than for other descriptor-based ML approaches (Liang *et al.*, 2020; Zhao *et al.*, 2020; Li *et al.*, 2021*a*). In the context of polymorphism, the predicted symmetry is expected to correspond to polymorph(s) (meta)stable at atmospheric conditions, since these are the standard circumstances for the entries in crystallographic databases.

4.3. Variable importance

The influence of each variable was evaluated from the decrease in accuracy upon its removal from the descriptor set.

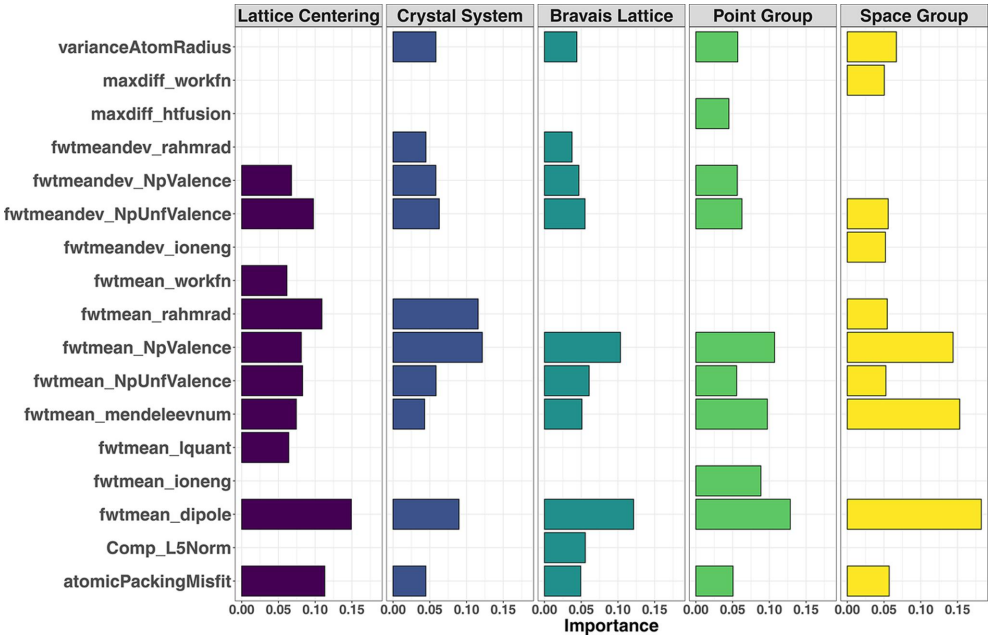


Figure 9
Variable importance in the RF models for lattice centering, crystal system, Bravais lattice, point group and space group. For brevity, only the ten most influential variables are shown for each symmetry category. The length of the bars is a quantitative measure of the decrease in accuracy upon removal of the variable from the set of descriptors. Gaps, *i.e.* absence, of a bar for a variable in a symmetry category indicate that the variable is not among the top-ten contributors. More information on each descriptor can be found in the SI.

Fig. 9 shows the variable importance for the five different symmetry categories (for brevity only the top-ten variables are shown). Several top-ranking variables are shared between the symmetry categories, albeit with varying impacts on the response as revealed by the different lengths of the bars, while some are specific to the symmetry category. Important descriptors include the fwtmean_dipole and the fwtmeandev_NpUnfValence that point to the weighted mean values of the atomic dipole and unfilled *p* orbitals, respectively. The atomicPackingMisfit is an indicator of the atomic packing efficiency (Guo *et al.*, 2011; Wang *et al.*, 2015). Another variable with a dominant influence on the model outcome is the Mendeleev number (mendeleevnum), which can be seen as a combination of important properties such as atomic size and electronegativity into a single parameter (Allahyari & Oganov, 2020). This information establishes the background for a fundamental definition of the filling rules for space groups in inorganic compounds.

5. Concluding remarks

The present work demonstrates that an ensemble decision-tree-based approach can achieve high accuracy in the symmetry prediction of new compounds using solely their elemental composition and the crystallographic information already available to science. Although class skewness is an intrinsic property of crystallographic databases, we have shown that reasonably accurate results can be achieved in space-group prediction if the classes used to train the models comprise more than 100 entries. Currently, the best ML approaches are limited to about 172 space groups with sufficient data out of the 230 classes. Therefore, specific efforts to populate the sparse classes must be made to fulfill a sound information goal and accelerate the discovery of materials with unusual space groups. Another critical aspect is the stoichiometric precision in chemical formulas, which requires suitable curation so that unique compounds can be accurately discriminated. On a final note, a meaningful contribution of ML to crystallography in the context of polymorphism will require the availability of significantly more data on polymorph structures, as well as suitably curated stability ranges in terms of temperature and pressure.

6. Related literature

The following references are cited in the SI: Chellali *et al.* (2019), Chen *et al.* (2020), Fu *et al.* (2021), Gao *et al.* (2018), Generalic (2020), Gild *et al.* (2016, 2019), Glawe *et al.* (2016), Gould & Bučko (2016), Guedri *et al.* (2021), Jadhav *et al.* (2021), Joseph *et al.* (2020), Jung *et al.* (2021), Lilensten *et al.* (2014), Liu *et al.* (2020, 2021*a*,*b*), KnowledgeDoor (2020), Manglam & Kar (2022), Marik *et al.* (2019), Mayandi *et al.* (2021, 2022), Motla *et al.* (2022), Nygard *et al.* (2020), Oses *et al.* (2020), Rahm *et al.* (2019), Rost *et al.* (2015), Sharma *et al.* (2018), Tekgül *et al.* (2022), Uporov *et al.* (2020), Witte *et al.* (2019), Wu *et al.* (2021, 2022), Yusenko *et al.* (2017), Zhu *et al.* (2020), Zlotea *et al.* (2019).

7. Data and software availability

The data sets used in this study are available from the corresponding public repositories – OQMD (https://oqmd.org),

Materials Project (https://materialsproject.org), COD (https://www.crystallography.net/cod). The ICSD and Pearson databases require commercial licenses. The prediction models have been added to an easy-to-use graphical user interface for public use. Instructions for software download and usage can be viewed at https://gitlab.com/vishsoft/cozy.

Funding information

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

References

Aguiar, J. A., Gong, M. L. & Tasdizen, T. (2020). *Comput. Mater. Sci.* **173**, 109409.
Allahyari, Z. & Oganov, A. R. (2020). *J. Phys. Chem. C*, **124**, 23867–23878.
Alsaui, A., Alqahtani, S. M., Mumtaz, F., Ibrahim, A. G., Mohammed, A., Muqaibel, A. H., Rashkeev, S. N., Baloch, A. A. B. & Alharbi, F. H. (2022). *Sci. Rep.* **12**, 1577.
Arik, S. O. & Pfister, T. (2019). *arXiv*:1908.07442.
Artstein, R. & Poesio, M. (2008). *Comput. Linguist.* **34**, 555–596.
ASM International (2021). *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds* (on DVD), release 2020/21. ASM International, Materials Park, Ohio, USA.
Axelrod, S., Schwalbe-Koda, D., Mohapatra, S., Damewood, J., Greenman, K. P. & Gómez-Bombarelli, R. (2022). *Acc. Mater. Res.* **3**, 343–357.
Chellali, M. R., Sarkar, A., Nandam, S. H., Bhattacharya, S. S., Breitung, B., Hahn, H. & Velasco, L. (2019). *Scr. Mater.* **166**, 58–63.
Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. (2019). *Chem. Mater.* **31**, 3564–3572.
Chen, Y., Xu, Z., Wang, M., Li, Y., Wu, C. & Yang, Y. (2020). *Mater. Sci. Eng. A*, **792**, 139774.
Corriero, N., Rizzi, R., Settembre, G., Del Buono, N. & Diacono, D. (2023). *J. Appl. Cryst.* **56**, 409–419.
Downs, R. T. & Hall-Wallace, M. (2003). *Am. Mineral.* **88**, 247–250.
Fu, M., Ma, X., Zhao, K., Li, X. & Su, D. (2021). *iScience*, **24**, 102177.
Gao, M. C., Miracle, D. B., Maurice, D., Yan, X., Zhang, Y. & Hawk, J. A. (2018). *J. Mater. Res.* **33**, 3138–3155.
Generalic, E. (2020). *Periodic Table of the Elements, Calculators, and Printable Materials*. https://www.periodni.com/.
Gild, J., Braun, J., Kaufmann, K., Marin, E., Harrington, T., Hopkins, P., Vecchio, K. & Luo, J. (2019). *J. Materiomics*, **5**, 337–343.
Gild, J., Zhang, Y., Harrington, T., Jiang, S., Hu, T., Quinn, M. C., Mellor, W. M., Zhou, N., Vecchio, K. & Luo, J. (2016). *Sci. Rep.* **6**, 37946.
Glawe, H., Sanna, A., Gross, E. K. U. & Marques, M. A. L. (2016). *New J. Phys.* **18**, 093011.
Goodall, R. E. A. & Lee, A. A. (2020). *Nat. Commun.* **11**, 6280.
Gould, T. & Bučko, T. (2016). *J. Chem. Theory Comput.* **12**, 3603–3613.
Grinsztajn, L., Oyallon, E. & Varoquaux, G. (2022). *Advances in Neural Information Processing Systems*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho & A. Oh, Vol. 35, pp. 507–520. New York: Curran Associates.
Guedri, A., Mnefgui, S., Hcini, S., Hlil, E. K. & Dhahri, A. (2021). *J. Solid State Chem.* **297**, 122046.
Guo, S., Ng, C., Lu, J. & Liu, C. T. (2011). *J. Appl. Phys.* **109**, 103505.
Hautier, G. (2019). *Comput. Mater. Sci.* **163**, 108–116.
Jadhav, M. S., Sahane, D., Verma, A. & Singh, S. (2021). *Adv. Powder Technol.* **32**, 378–384.
Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G. & Persson, K. A. (2013). *APL Mater.* **1**, 011002.
Jha, D., Ward, L., Paul, A., Liao, W., Choudhary, A., Wolverton, C. & Agrawal, A. (2018). *Sci. Rep.* **8**, 17593.
Joseph, J., Haghdadi, N., Annasamy, M., Kada, S., Hodgson, P. D., Barnett, M. R. & Fabijanic, D. M. (2020). *Scr. Mater.* **186**, 230–235.
Jung, Y., Lee, K., Hong, S. J., Lee, J. K., Han, J., Kim, K. B., Liaw, P. K., Lee, C. & Song, G. (2021). *J. Alloys Compd.* **886**, 161187.
Kadra, A., Lindauer, M., Hutter, F. & Grabocka, J. (2021). *Thirty-Fifth Conference on Neural Information Processing Systems*. New York: Curren Associates.
KnowledgeDoor (2020). https://www.knowledgedoor.com.
Kong, S., Guevarra, D., Gomes, C. P. & Gregoire, J. M. (2021). *Appl. Phys. Rev.* **8**, 021409.
Kusaba, M., Liu, C. & Yoshida, R. (2022). *Comput. Mater. Sci.* **211**, 111496.
Li, Y., Dong, R., Yang, W. & Hu, J. (2021*a*). *Comput. Mater. Sci.* **198**, 110686.
Li, Y., Yang, W., Dong, R. & Hu, J. (2021*b*). *ACS Omega*, **6**, 11585–11594.
Liang, H., Stanev, V., Kusne, A. G. & Takeuchi, I. (2020). *Phys. Rev. Mater.* **4**, 123802.
Lilensten, L., Couzinié, J. P., Perrière, L., Bourgon, J., Emery, N. & Guillot, I. (2014). *Mater. Lett.* **132**, 123–125.
Liu, B., Wu, J., Cui, Y., Zhu, Q., Xiao, G., Wu, S., Cao, G. & Ren, Z. (2020). *Scr. Mater.* **182**, 109–113.
Liu, B., Wu, J., Cui, Y., Zhu, Q., Xiao, G., Wu, S., Cao, G. & Ren, Z. (2021*a*). *J. Alloys Compd.* **869**, 159293.
Liu, C.-H., Tao, Y., Hsu, D., Du, Q. & Billinge, S. J. L. (2019). *Acta Cryst.* A**75**, 633–643.
Liu, J., Xu, J., Sleiman, S., Chen, X., Zhu, S., Cheng, H. & Huot, J. (2021*b*). *Int. J. Hydrogen Energy*, **46**, 28709–28718.
Manglam, M. K. & Kar, M. (2022). *J. Alloys Compd.* **899**, 163367.
Marik, S., Motla, K., Varghese, M., Sajilesh, K. P., Singh, D., Breard, Y., Boullay, P. & Singh, R. P. (2019). *Phys. Rev. Mater.* **3**, 060602.
Marzari, N., Ferretti, A. & Wolverton, C. (2021). *Nat. Mater.* **20**, 736–749.
Mayandi, J., Dias, M., Stange, M., Lind, A., Sunding, M. F., Cerdeira, A. C., Schrade, M., Belle, B. D., Finstad, T. G., Pereira, L. C. J., Diplas, S. & Carvalho, P. A. (2021). *Materialia*, **20**, 101250
Mayandi, J., Finstad, T. G., Dahl, Ø., Vajeeston, P., Schrade, M., Løvvik, O. M., Diplas, S. & Carvalho, P. A. (2022). *Thin Solid Films*, **744**, 139083.
Motla, K., Soni, V., Meena, P. K. & Singh, R. P. (2022). *Supercond. Sci. Technol.* **35**, 074002.
Nygård, M. M., Sławiński, W. A., Ek, G., Sørby, M. H., Sahlberg, M., Keen, D. A. & Hauback, B. C. (2020). *Acta Mater.* **199**, 504–513.
Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. (2019). *Nat. Rev. Mater.* **4**, 331–348.
Oses, C., Toher, C. & Curtarolo, S. (2020). *Nat. Rev. Mater.* **5**, 295–309.
Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N. T. P., Ramasamy, S., DeCost, B. L., Tian, S. I. P., Romano, G., Kusne, A. G. & Buonassisi, T. (2019). *npj Comput. Mater.* **5**, 60.
Park, W. B., Chung, J., Jung, J., Sohn, K., Singh, S. P., Pyo, M., Shin, N. & Sohn, K.-S. (2017). *IUCrJ*, **4**, 486–494.
Rahm, M., Zeng, T. & Hoffmann, R. (2019). *J. Am. Chem. Soc.* **141**, 342–351.
Rost, C. M., Sachet, E., Borman, T., Moballegh, A., Dickey, E. C., Hou, D., Jones, J. L., Curtarolo, S. & Maria, J.-P. (2015). *Nat. Commun.* **6**, 8485.
Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. (2013). *JOM*, **65**, 1501–1509.
Saal, J. E., Oliynyk, A. O. & Meredig, B. (2020). *Annu. Rev. Mater. Res.* **50**, 49–69.
Sharma, Y., Musico, B. L., Gao, X., Hua, C., May, A. F., Herklotz, A., Rastogi, A., Mandrus, D., Yan, J., Lee, H. N., Chisholm, M. F., Keppens, V. & Ward, T. Z. (2018). *Phys. Rev. Mater.* **2**, 060404.

Shwartz-Ziv, R. & Armon, A. (2022). *Information Fusion*, **81**, 84–90.
Sun, L., Zhou, Y.-X., Wang, X.-D., Chen, Y.-H., Deringer, V. L., Mazzarello, R. & Zhang, W. (2021). *npj Comput. Mater.* **7**, 29.
Suzuki, Y., Hino, H., Hawai, T., Saito, K., Kotsugi, M. & Ono, K. (2020). *Sci. Rep.* **10**, 21790.
Tekgül, A., Sarlar, K., Küçük, N. & Etemoğlu, A. B. (2022). *Phys. Scr.* **97**, 075814.
Uporov, S. A., Ryltsev, R. E., Bykov, V. A., Estemirova, S. K. & Zamyatin, D. A. (2020). *J. Alloys Compd.* **820**, 153228.
Urusov, V. & Nadezhina, T. (2009). *J. Struct. Chem.* **50**(Suppl. 1), 22–37.
Vaitkus, A., Merkys, A. & Gražulis, S. (2021). *J. Appl. Cryst.* **54**, 661–672.
Vargaftik, S. & Ben-Itzhak, Y. (2022). *Proceedings of the 2nd European Workshop on Machine Learning and Systems*, pp. 10–19. ACM.
Venkatraman, V. (2021). *Comput. Mater. Sci.* **197**, 110637.
Venkatraman, V. (2023). *Front. Chem.* **11**, 1239467.
Venkatraman, V. & Carvalho, P. A. (2022). *Acta Mater.* **240**, 118353.
Wang, Z., Huang, Y., Yang, Y., Wang, J. & Liu, C. (2015). *Scr. Mater.* **94**, 28–31.
Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. (2016). *npj Comput. Mater.* **2**, 16028.
Witte, R., Sarkar, A., Kruk, R., Eggert, B., Brand, R. A., Wende, H. & Hahn, H. (2019). *Phys. Rev. Mater.* **3**, 034406.
Wu, S., Qiao, D., Zhang, H., Miao, J., Zhao, H., Wang, J., Lu, Y., Wang, T. & Li, T. (2022). *J. Mater. Sci. Technol.* **97**, 229–238.
Wu, S., Qiao, D., Zhao, H., Wang, J. & Lu, Y. (2021). *J. Alloys Compd.* **889**, 161800.
Yusenko, K. V., Riva, S., Carvalho, P. A., Yusenko, M. V., Arnaboldi, S., Sukhikh, A. S., Hanfland, M. & Gromilov, S. A. (2017). *Scr. Mater.* **138**, 22–27.
Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. (2019). *J. Appl. Cryst.* **52**, 918–925.
Zhao, Y., Cui, Y., Xiong, Z., Jin, J., Liu, Z., Dong, R. & Hu, J. (2020). *ACS Omega*, **5**, 3596–3606.
Zhu, S., Chen, X., Liu, J., Yang, N., Chen, J., Gu, C., Cheng, H., Yan, K., Zhu, Z. & Wang, K. (2020). *Mater. Sci. Eng. B*, **262**, 114777.
Zlotea, C., Sow, M. A., Ek, G., Couzinié, J., Perrière, L., Guillot, I., Bourgon, J., Møller, K. T., Jensen, T. R., Akiba, E. & Sahlberg, M. (2019). *J. Alloys Compd.* **775**, 667–674.