Machine learning to predict quasicrystals from chemical compositions

Chang Liu
The Institute Of Statistical Mathematics https://orcid.org/0000-0002-9511-4283
Erina Fujita
The University of Tokyo
Yukari Katsura
The University of Tokyo
Yuki Inada
The University of Tokyo
Asuka Ishikawa
Tokyo University of Science
Ryuji Tamura
Tokyo University of Science
Kaoru Kimura
University of Tokyo
Ryo Yoshida (yoshidar@ism.ac.jp)
The Institute of Statistical Mathematics https://orcid.org/0000-0001-8092-0162

Article

Keywords: Quasicrystal, Approximant, Machine learning, Materials informatics, High-throughput screen32 ing

Posted Date: February 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-240290/v1

Version of Record: A version of this preprint was published at Advanced Materials on July 19th, 2021. See the published version at https://doi.org/10.1002/adma.202102507.

Machine learning to predict quasicrystals from chemical compositions

Chang Liu[1], Erina Fujita[2], Yukari Katsura[2], Yuki Inada[2], Asuka Ishikawa[3], Ryuji Tamura[3], Kaoru Kimura[2,*], and Ryo Yoshida[1,4,5,*]

[1]Research Organization of Information and Systems, The Institute of Statistical Mathematics, Tachikawa, 190-8562, Japan
[2]The University of Tokyo, Department of Advanced Materials Science, Kashiwa, 277-8561, Japan
[3]Tokyo University of Science, Department of Materials Science and Technology, Tokyo, 125–8585, Japan
[4]National Institute for Materials Science, Research and Service Division of Materials Data and Integrated System, Tsukuba, 305-0047, Japan
[5]The Graduate University for Advanced Studies, Department of Statistical Science, Tachikawa, 190-8562, Japan
[*]yoshidar@ism.ac.jp, bkimura@phys.mm.t.u-tokyo.ac.jp

Abstract

Quasicrystals have emerged as a new class of solid-state materials that have long-range order without periodicity, exhibiting rotational symmetries that are disallowed for periodic crystals in most cases. To date, hundreds of new quasicrystals have been found, leading to the discovery of many new and exciting phenomena. However, the pace of the discovery of new quasicrystals has slowed in recent years, largely owing to the lack of clear guiding principles for the synthesis of new quasicrystals. Here, we show that the discovery of new quasicrystals can be accelerated with a simple machine learning workflow. With a list of the chemical compositions of known quasicrystals, approximant crystals, and ordinary crystals, we trained a prediction model to solve the three-class classification task and evaluated its predictability compared to the observed phase diagrams of ternary aluminum systems. The validation experiments strongly support the superior predictive power of machine learning, with the precision and recall of the phase prediction task reaching approximately 0.793 and 0.714, respectively. Furthermore, analyzing the input–output relationships black-boxed into the model, we identified nontrivial empirical equations interpretable by humans that describe conditions necessary for quasicrystal formation.

Keywords: Quasicrystal, Approximant, Machine learning, Materials informatics, High-throughput screening

Introduction

This study demonstrates the potential of machine learning to predict quasicrystal compositions. Quasicrystals do not have the translational symmetry of ordinary crystals but have a high degree of order in their atomic arrangement. The first quasicrystal was discovered by Schechtman in 1984 [1]. A few years later, Tsai and his colleagues discovered a series of stable quasicrystals in systems including Al-Cu-Fe, Al-Ni-Co, Al-Pd-Mn, Yb-Cd, and Yb–Cd–Mg [2–6]. Since then, 100 or so new quasicrystals have been discovered. In the history of quasicrystal research, the discovery of new quasicrystals has unearthed new and interesting phenomena such as anomalous electronic properties [7, 8], insulating behaviors [9], valence fluctuation [10], quantum criticality [11], superconductivity [12], and so on. However, the pace of the discovery of new quasicrystals has slowed significantly in recent years. Figure 1 (a) shows the annual trend of new quasicrystals found in aluminum alloy systems. From 1984 to 1999, new quasicrystals were discovered at a rate of about two per year. On the other hand, in recent

years, the frequency of new discoveries has dramatically decreased. This recent trend is mainly due to the fact that no clear guiding principles have been established for the synthesis of new quasicrystals. In terms of the stability mechanism of quasicrystals, the Hume-Rothery rules [13], i.e. itinerant valence electron concentration, e/a, and atomic size factor, have been considered [14, 15]. Thus, this study aimed to accelerate the discovery of new quasicrystals by introducing machine learning to the field.

Recently, a wide variety of machine learning technologies has been rapidly introduced to materials science. In particular, high-throughput screening (HTS) across extensive libraries of candidate materials, which typically contain millions or even billions of virtually created candidates, is a promising machine learning application. HTS relies on the fast computation of a statistical model that describes physical, chemical, electronic, thermodynamic, and mechanical properties and unobserved structural features as a function of the material. Nowadays, many successful case studies of HTS have been reported. The range of applications is broad, including small organic molecules [16–18], polymeric materials [19], and inorganic solid-state materials [20–23]. Can HTS based on machine learning also contribute to the discovery of quasicrystals? We seek to answer this question.

The analytical workflow of this study consisted of simple supervised learning. The input variable of the model is a chemical composition, which is characterized by a descriptor vector of 232 compositional features. The output variable is a class label corresponding to one of three structural categories: quasicrystal (QC), approximant crystal (AC), and "others," which includes ordinary crystals. A list of the chemical compositions of known quasicrystals, approximants, and ordinary crystals was used as the training data. We systematically evaluated the potential predictability of the proposed machine learning model for the three-class classification problem. Furthermore, virtual screening of all ternary alloy systems containing aluminum and transition elements was conducted for the entire search space. The phase prediction results were compared with 30 experimental phase diagrams extracted from the literature, and the predictability was investigated in detail. The precision and recall rates for the phase prediction task reached approximately 0.793 and 0.714, respectively. Furthermore, by revealing the input–output landscape inherently encoded in the black-box model, we identified the law of compositional features relevant to the formation of quasicrystalline and approximant crystalline phases. This rule of thumb could be expressed by simple mathematical equations describing a set of compositional features such as the distribution of van der Waals radii of atoms and valence electron concentration. With this study, we take the first step toward enabling the data-driven discovery of innovative quasicrystals.

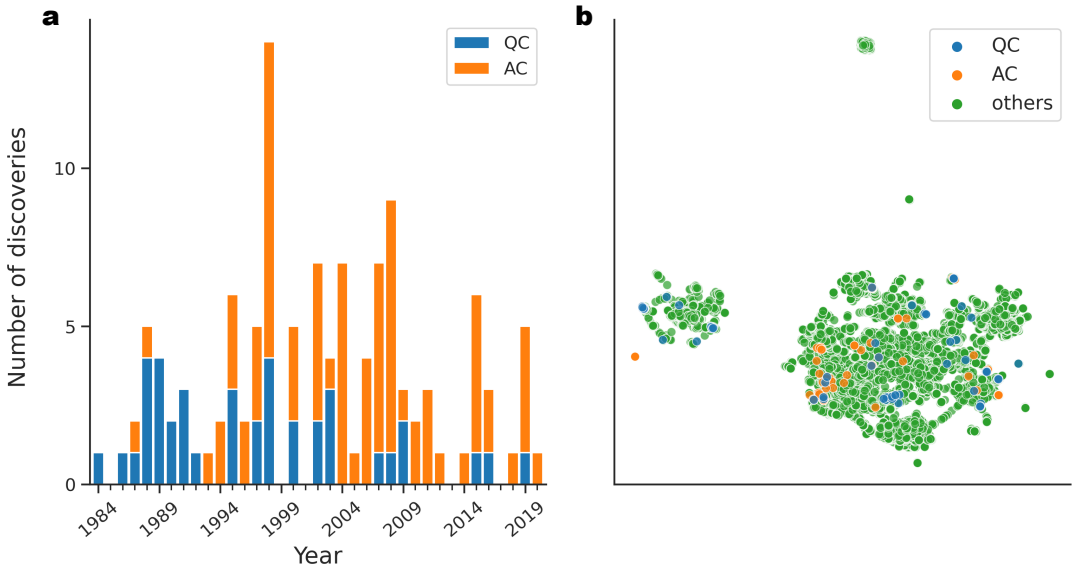

Figure 1: Quasicrystals (QC) and approximants (AC) that have been discovered so far. **a.** Annual trend in the discovery of new quasicrystals (blue) and approximant crystals (orange) in aluminum alloys. **b.** Distribution of the compositional dataset that was visualized onto a two-dimensional space obtained by the UMAP algorithm [24] (see the Methods section). Quasicrystals, approximants, and ordinary crystals are color-coded by blue, orange, and green, respectively.

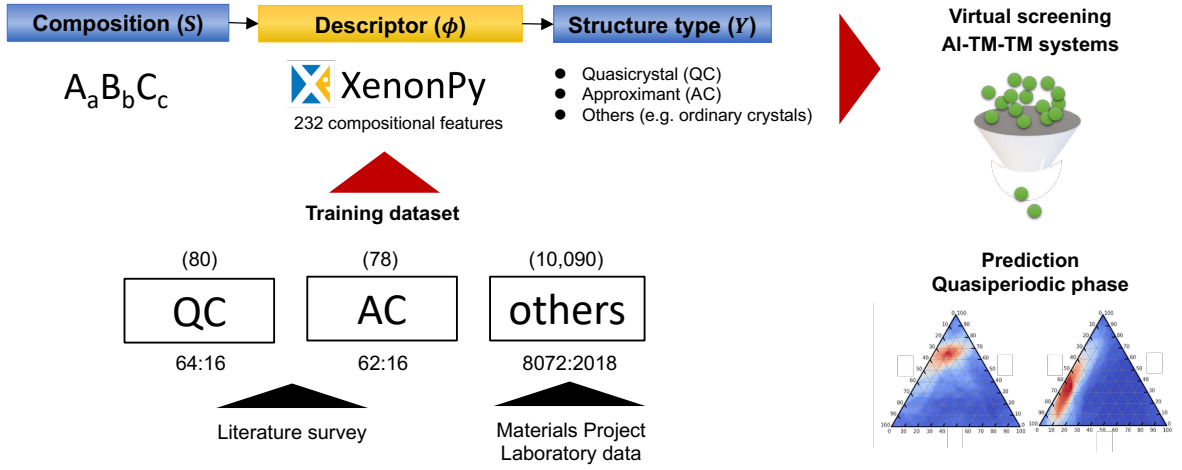

Figure 2: Machine learning workflow. The compositional features were encoded into a 232-dimensional descriptor vector, and a prediction model was created using a random forest classifier. The trained model predicts the class label of a given chemical composition as quasicrystal (QC), approximant (AC), or others. Model training and testing were performed on the compositional features of 80 known quasicrystals, 78 approximants, and 10,090 ordinary crystals. Finally, we performed HTS across all Al-TM-TM (TM: transition metal) alloys to generate their predicted phase diagrams. The results were compared with experimental phase diagrams obtained from the literature.

Table 1: Prediction performance for the three-class classification problem of quasicrystals (QC), approximants (AC), and others. The left table is the confusion matrix, and the right table reports the per-class recall, precision, and F_1 metrics. The performance metrics were averaged over 100 different bootstrap sets, and the numbers in parentheses represent the standard deviations.

		Predicted class		
		QC	AC	others
True class	QC	9.63 (1.641)	3.24 (1.342)	3.13 (1.189)
	AC	3.11 (1.555)	9.73 (1.805)	3.16 (1.573)
	others	0.76 (0.896)	0.42 (0.619)	2016.82 (1.024)

	Recall	Precision	F_1
QC	0.602 (0.103)	0.722 (0.090)	0.650 (0.076)
AC	0.608 (0.113)	0.731 (0.089)	0.658 (0.088)
others	0.999 (0.001)	0.997 (0.001)	0.998 (0.001)

Results

Machine learning workflow

We used a set of chemical compositions and their class labels for model training. The class labels were QC, AC, and "others" representing types other than the first two. We compiled a list of 80 quasicrystals and 78 approximants from the Crystallography of Quasicrystals handbook [25] (see Table S1 in the Supplementary Note and Supplementary Data for digital data). In addition, the compositions of 10,000 ordinary crystals were randomly extracted from the Materials Project database [26] and 90 from our laboratory data on failed quasicrystal syntheses. These instances form the class "others". The detailed data preparation procedure is given in the Methods section.

The machine learning workflow is summarized in Figure 2. The features of a given composition were encoded into a descriptor vector of length 232. The details of the compositional descriptor are described later. The model describes the class label as a function of the descriptor vector of a given composition. We built various models with random forests and neural networks, but since there was no significant difference in prediction performance, this paper presents only the former results. The model training procedure is detailed in the Methods section.

For each class, approximately 80% of the total data was randomly selected for training (64, 62, and 8,072 for QC, AC, and others, respectively), and the remaining were used as a test set to measure the prediction performance (16, 16, and 2,018 for QC, AC, and others, respectively). The configuration of hyperparameters was selected so as to optimize the overall prediction accuracy in the cross validation that was looped within

the training set (for the list of hyperparameters and their search range, see the Methods section). To mitigate the effect of sampling bias on the assessment of predictive performance, we performed 100 random splits of the training and test sets and calculated the mean and variance of the resulting performance metrics.

Representation of compositional features

Here, we describe the compositional descriptor. The chemical formula is denoted by $S = S_{c^1}^1 S_{c^2}^2 \cdots S_{c^K}^K$. Each element of the descriptor vector of length 232 takes the form

$$\phi_{f,\eta}(S) = f(c^1, \ldots, c^K, \eta(S^1), \ldots, \eta(S^K)).$$

The notation $\eta(S^k)$ on the right-hand side denotes a feature quantity of element S^k, such as the atomic weight, electronegativity, or polarizability. With the function f, the K element features $\eta(S^1), \ldots, \eta(S^K)$ with fraction $c^1, \ldots, c^K$ were converted into compositional features. For f, we operated with the weighted average, weighted variance, max-pooling, and min-pooling as given by

$$\begin{aligned}
\phi_{\mathrm{ave},\eta}(S) &= \frac{1}{\sum_{k=1}^K c^k} \sum_{k=1}^K c^k \eta(S^k), \\
\phi_{\mathrm{var},\eta}(S) &= \frac{1}{\sum_{k=1}^K c^k} \sum_{k=1}^K c^k (\eta(S^k) - \phi_{\mathrm{ave},\eta}(S))^2, \\
\phi_{\mathrm{max},\eta}(S) &= \max\{\eta(S^1), \ldots, \eta(S^K)\}, \\
\phi_{\mathrm{min},\eta}(S) &= \min\{\eta(S^1), \ldots, \eta(S^K)\}.
\end{aligned}$$

Table S3 in the Supplementary Note provides a list of the 58 element features that were implemented in XenonPy, a Python open-source platform for materials informatics that we developed [27]. The element feature set includes the atomic number, bond radius, van der Waals radius, electronegativity, thermal conductivity, bandgap, polarizability, boiling point, melting point, number of valence electrons in each orbital, and so on.

Generalization ability of the model

We predicted the class labels of 2,050 test compositions with the 100 trained models. The confusion matrix shown in Table 1 and resulting performance metrics suggest that the machine learning models were successful in gaining predictive capability. In this analysis, we examined the prediction performance based on three metrics: recall, precision, and F-value. These metrics quantified the predictive performance for each class c of QC, AC, and others according to

$$\begin{aligned}
\mathrm{Recall}(c) &= \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FN}(c)}, \\
\mathrm{Precision}(c) &= \frac{\mathrm{TP}(c)}{\mathrm{TP}(c) + \mathrm{FP}(c)}, \\
\mathrm{F}_1(c) &= 2 \cdot \frac{\mathrm{Recall}(c) \cdot \mathrm{Precision}(c)}{\mathrm{Recall}(c) + \mathrm{Precision}(c)}.
\end{aligned}$$

$\mathrm{TP}(c)$ denotes the number of true positives when label c is treated as positive and the other two classes as negative, and $\mathrm{FN}(c)$ and $\mathrm{FP}(c)$ represent a false negative and false positive, respectively. Thus, the recall rate represents the fraction of compositions with true class label c that could be predicted as c, whereas the precision represents the fraction of compositions predicted as label c that were actually label c. There is a tradeoff between the recall and precision rates. $\mathrm{F}_1(c)$ is the harmonic mean of the recall and precision.

The precision and recall for the prediction of the class "others" reached 0.997 and 0.999, respectively. This means that almost perfect predictions were achieved for the binary classification of QC/AC as a merged class versus others. On the other hand, the precision and recall were 0.722 and 0.602 for QC and 0.731 and 0.608 for AC, respectively. Although the classification performance was slightly lower than that in the prediction of the class "others", the trained models exhibits the generalized ability to identify chemical compositions that could potentially generate quasicrystals and approximant crystals.

Phase prediction of ternary alloy systems

High-throughput virtual screening of all composition spaces was performed on a total of 1,080 systems of Al-TM[4,5]-TM[4,5] and Al-TM[4,5]-TM[6], where the numbers in square brackets denote the periods of the transition elements. In addition, we added a set of non-transition-metal elements {Mg, Si, Ga, Ge, In, Sn, Sb} in place of

TM[4,5] and {Tl, Pb, Bi} in place of TM[6]. With a given model, the class probability of QC, AC, or others was calculated for a given chemical composition. For each composition, we standardized its fractions into relative proportions. A ternary phase diagram was gridded with 20,301 points by dividing the interval of the composition ratio from 0 to 1 by 200 equally spaced grid points. A label exhibiting the maximum probability was assigned to each grid point in the diagram. In this way, quasicrystalline and approximant phases were predicted. Using this screening process, quasicrystalline phases were predicted to exist in 185 systems, which would be an overestimate. Notably, in 136 of the 185 systems, the predicted quasicrystalline and approximant phases coexisted in neighboring regions of the same diagram. This result is highly consistent with experimental observations, which we give examples of later.

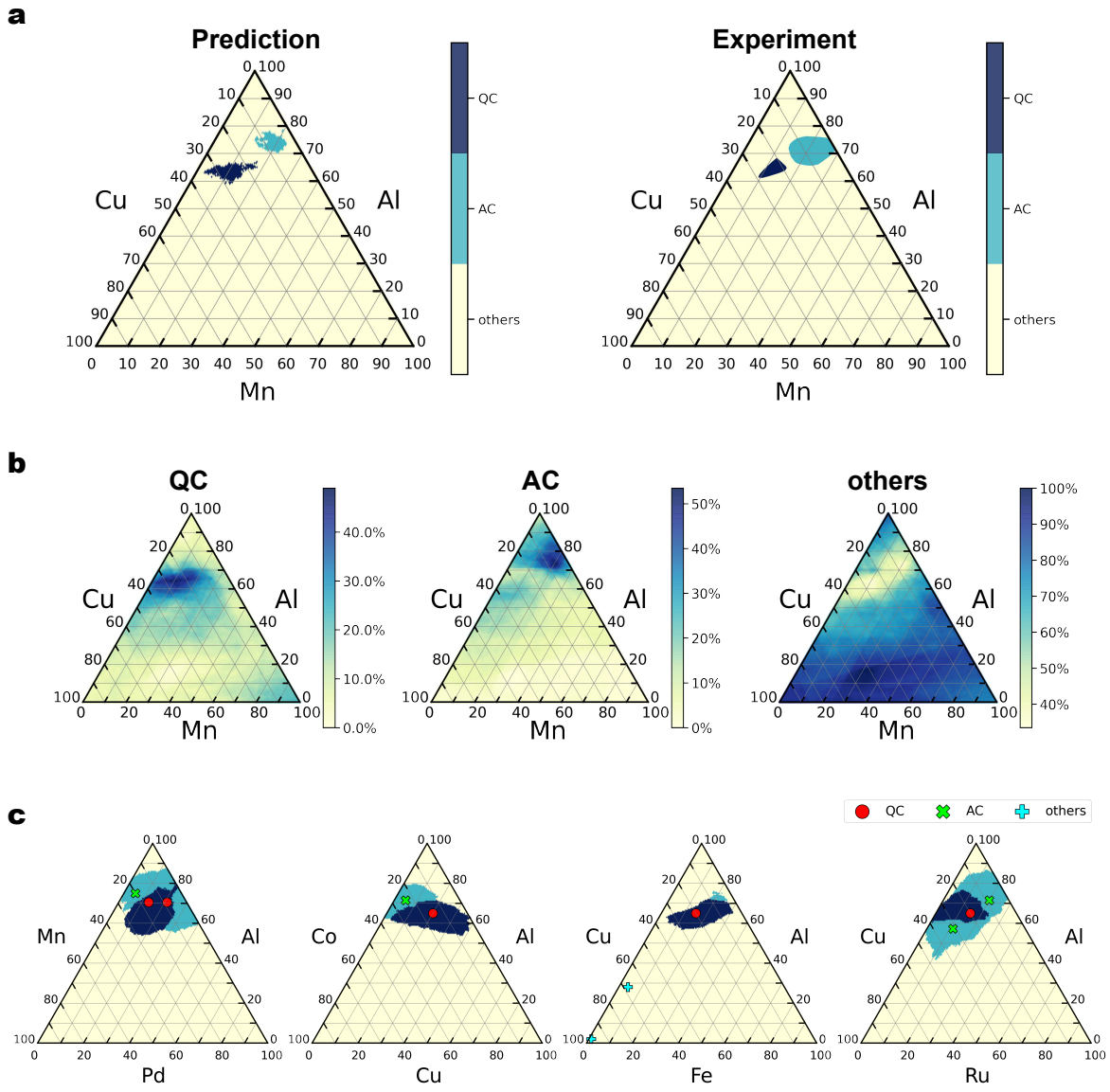

Figure 3: Phase prediction of the Al-Mn-Cu system. **a.** Predicted phase diagram (left panel) and experimental phase diagram (right panel) of the Al-Mn-Cu system. The three colors denote the quasicrystalline phase (QC), approximant phase (AC), and others. Despite the lack of training instances for Al-Mn-Cu, the model successfully predicts the unseen quasicrystalline and approximant crystalline phases. **b.** Heatmap display of the predicted class probability of QC, AC, and others for the Al-Mn-Cu system. **c.** In order to observe the training instances relevant to the model decision making, we examined the distribution of training instances in the four ternary systems closest to Al-Mn-Cu.

We verified the validity of the predicted phase diagrams based on the experimental quasicrystal and approximant phase regions of the 30 systems that were extracted from the literature [28–52]. We found 198 papers

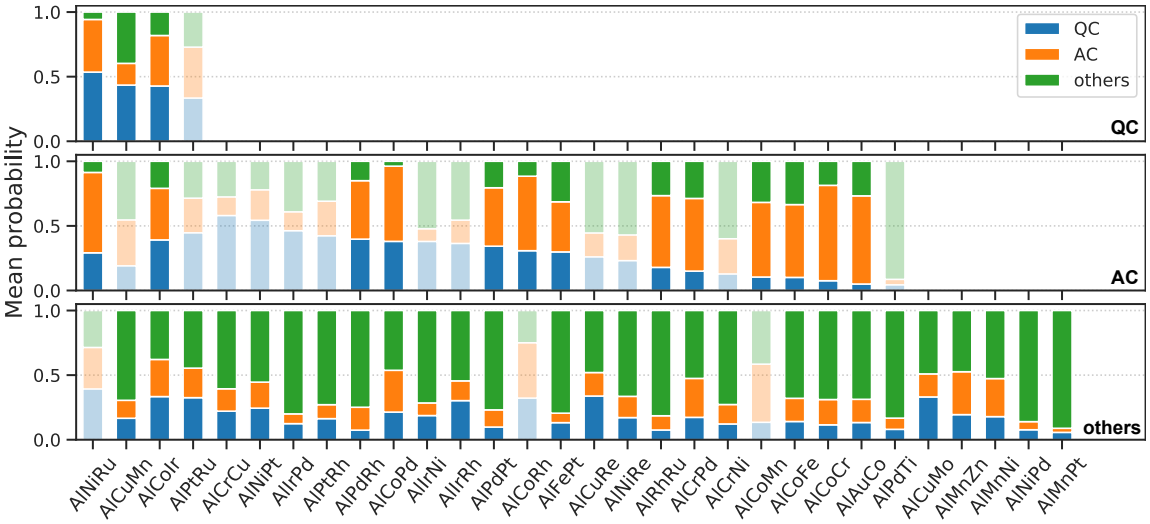

Figure 4: Prediction performance for the 30 different Al-TM-TM systems. The mean class probability was calculated in each of the experimental phase regions (top: QC, middle: AC, and bottom: others) using our trained random forest classifier. The bar plots shown in a transparent color represent phases where class label prediction based on maximum probability failed.

published by Prof. Grushko's group, which include ternary phase diagrams of Al-transition elements encompassing 64 unique alloy systems. Excluding the systems containing the 80 quasicrystal and 78 approximant compositions used for training, the remaining 30 systems were used for performance evaluation. Figure S2 in the Supplementary Note displays all the predicted and experimental phase diagrams, and Figure 3 shows an example. With a given classifier, the class probability of forming quasicrystals, approximants, or others was drawn on the phase diagram of Al-Cu-Mn [46]. To evaluate the prediction performance, the agreement between the three class probabilities and the experimental quasicrystalline and approximant phase regions was investigated. For each ternary system, $\mathcal{G}_{c_{\mathrm{exp}}}$ denotes the set of all grid points in experimental phase $c_{\mathrm{exp}} \in \{\mathrm{QC}, \mathrm{AC}, \mathrm{others}\}$ in a diagram. Using the trained model, we calculated the mean probability $p(Y = c|\mathcal{G}_{c_{\mathrm{exp}}})$ for each c_{exp} and $c \in \{\mathrm{QC}, \mathrm{AC}, \mathrm{others}\}$ by

$$p(Y = c|\mathcal{G}_{c_{\mathrm{exp}}}) = \frac{1}{|\mathcal{G}_{c_{\mathrm{exp}}}|} \sum_{i \in \mathcal{G}_{c_{\mathrm{exp}}}} p(Y_i = c|S_i).$$

$p(Y_i = c|S_i)$ denotes the predicted probability that class label Y_i of composition S_i with $i \in \mathcal{G}_{c_{\mathrm{exp}}}$ is equal to c. The probability values were averaged over a given phase with grid points $i \in \mathcal{G}_{c_{\mathrm{expt}}}$. If $p(Y = c|\mathcal{G}_{c_{\mathrm{expt}}})$ reaches a maximum at $c = c_{\mathrm{exp}}$, the prediction is correct. The prediction performance across the 30 alloy systems is summarized in Table 2. In addition, the mean probability of each class with respect to the three different phases in the 30 systems is displayed in Figure 4.

Table 2: Phase prediction performance for the 30 Al-TM-TM alloy systems.

		Predicted class		
		QC	AC	others
True class	QC	3	1	0
	AC	5	13	7
	others	1	2	27

	Recall	Precision	F_1
QC	0.750	0.333	0.462
AC	0.520	0.813	0.634
others	0.900	0.794	0.844

The results showed a similar trend to the performance metrics from the previously discussed composition-level evaluation (Table 1). The precision and recall were 0.750 and 0.333 for quasicrystals, 0.520 and 0.813 for approximants, and 0.794 and 0.900 for others, respectively. The total precision and recall (micro-average) across the three-class classification reached 0.793 and 0.714, respectively. As shown in Figure 4, the number of cases in which a quasicrystalline phase region was misclassified as an approximant was 1/4, and the number of cases in which an approximant phase was misclassified as a quasicrystal was 5/25. On the other hand, the other regions, including the ordinary crystalline phases, were almost completely predictable. Although the misclassification rate

for quasicrystalline and approximant phases increased slightly, the trained model was found to have sufficient predictive power to be useful.

Although the misclassification rate between quasicrystalline and approximant phases was slightly high, we concluded that the model is more or less capable of identifying compositional regions of quasicrystals and approximant crystals. As illustrated in Figure 3 showing the Al-Mn-Cu phase diagram and its prediction results, in many cases, the model adequately captured not only the positional features of the quasicrystalline and approximant phases but also their contour shapes (see also Figure S1 in the Supplementary Note for all results). Interestingly, despite the lack of any training instances from the Al-Mn-Cu system, the model successfully predicted the two true phase regions. In order to identify the instances in the dataset on which the model relied in the training process, four other systems with the closest compositional patterns to Al-Mn-Cu were selected, and the distribution of the training data was examined (Figure 3 (c)). The compositional closeness was evaluated based on the Euclidean distance of the normalized 232-dimensional compositional descriptor. Simple pattern matching based on the similarity of the input and output to the training data never predicted the positional and geometric features of the quasicrystalline and approximant phases in the Al-Mn-Cu phase diagram. Thus, the model involves a higher-order recognition mechanism than simple nearest-neighbor matching.

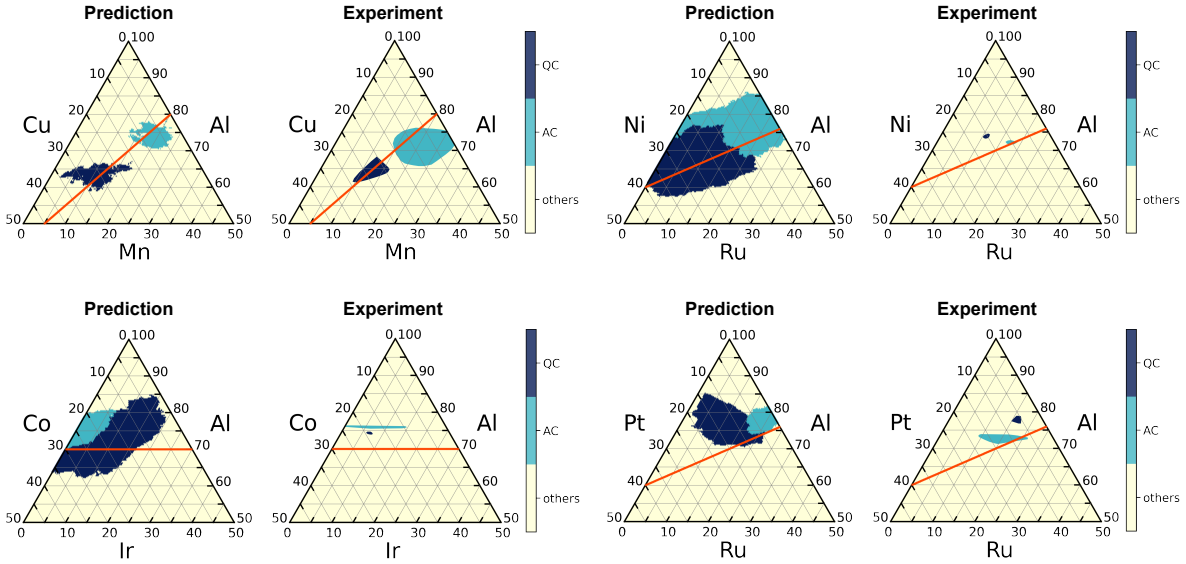

Figure 5: Predicted and experimental phase diagrams of four ternary alloy systems. The orange lines indicate the Hume-Rothery rule of valence electron concentration with $e/a = 1.8$.

Hume-Rothery's law autonomously learned by machine learning

Notably, it was found that the trained models learned Hume-Rothery's electron concentration law [13], which is one of the most widely applied empirical rules regarding the formation of stable quasicrystalline alloys. In 1990, Tsai et al. discovered a series of thermally stable quasicrystals in the Al-Cu-TM and Al-Pd-TM systems [2–4]. In a subsequent study, the discovered quasicrystals were found to obey Hume-Rothery's electron concentration law on the average itinerant valence electron number e/a [14].

Figure 5 shows the predicted and experimental phase diagrams for four of the 30 evaluated alloy systems as discussed above. In each diagram, the line where the average itinerant valence electron number follows $e/a = 1.80$ is overlaid (see [53] for details on the calculation of e/a). Surprisingly, in all the systems, the straight lines overlap with the predicted and true regions of quasicrystals and approximant crystals. In the 30 ternary alloy systems discussed above, the straight line completely overlapped with the predicted regions in 26 systems (Figure S1 in the Supplementary Note). Note that our compositional descriptors do not include e/a values; this widely known empirical rule occurred via the nonlinear mapping of our descriptors. If we can comprehensively extract such implicit rules inherent to the trained machine learning model, we could obtain hypothetical insights on the formation rules or mechanisms of quasicrystalline phases.

Why can the model predict quasicrystals?

To determine on what basis the model determines structural classes, we analyzed the predicted class labels $\{Y_i|i = 1, \dots N\}$ in relation to the 21,925,080 hypothetical compositions $\{S_i|i = 1, \dots, N\}$ $(N = 21,925,080)$ that were used in the HTS of the entire composition space of the 30 aluminum alloys. The model determined mathematical map $Y = f(S)$ between predicted label Y and descriptor vector $\phi(S) = (\phi_1(S), \dots, \phi_{232}(S)) \in \mathbb{R}^{232}$. First, we examined the degree of relevance of each descriptor element $\phi_h(S)$ $(h = 1, \dots, 232)$ with respect to predicted Y. As a quantitative measure of relevance, we applied the maximal information coefficient (MIC), a widely used measure of statistical independence (linear and nonlinear correlation) between two variables [54]. Using the dataset $\{(Y_i, \phi_h(S_i))|i = 1, \dots, N\}$, which was produced from the black-box machine learning model, we estimated the joint distribution $P(Y, \phi_h(S))$ and marginal distribution $Q(Y, \phi_h(S)) = P(Y)P(\phi_h(S))$, where the latter assumes independence between Y and $\phi_h(S)$. The MIC evaluates the statistical independence of the hth descriptor $\phi_h(S)$ and output Y by measuring the discrepancy between $P(Y, \phi_h(S))$ and $Q(Y, \phi_h(S))$. The Kullback–Leibler divergence, which is equivalent to the mutual information between $\phi_h(S)$ and Y, was employed for the MIC evaluation, and an adaptive binning algorithm was applied to approximate the two probability distributions by generating histograms.

Table 3 shows the top 20 most relevant descriptors as examples, which suggest that the weighted averages of the van der Waals radius, electronegativity, and first ionization energy are highly relevant to the basis of the model decision making process. The most relevant descriptor, i.e., the weighted average of the van der Waals radius, is consistent with the Hume-Rothery rules, where the atomic size factor is considered to contribute to the stability mechanism of icosahedral quasicrystals. In addition, Table 3 shows the within-class mean and within-class variance of the subset of $\{\phi_h(S_i)|i = 1, \dots, N\}$ belonging to each {QC, AC, others}. Descriptors with larger discrepancies in the within-class means and smaller discrepancies in the within-class variances are interpreted as having a high degree of separation between classes and thus a high degree of relevance to the output class label. Most of the listed relevant descriptors exhibited significantly large between-class separations in terms of QC/AC versus others or QC versus AC.

Only listing highly relevant descriptors is not enough to clarify the basis of the model decision making process. Instead, we want to derive an explicit empirical equation, such as the rule of $e/a = 1.8$ for itinerant valence electron concentration. In this study, we focused on the binary classification task of discriminating between merged QC/AC and others. We calculated the within-class mean m_h for the QC/AC group from the observed $\{\phi_h(S_i)|i = 1, \dots, N\}$ with their predicted $Y =$ QC or AC. It is expected that the model places a high classification probability $(Y \in \{\text{QC, AC}\}|S)$ on any composition ratio S that satisfies exactly or approximately $\phi_h(S) = m_h$. For example, in the case where S is a ternary system $S^1_{\hat{c}^1} S^2_{\hat{c}^2} S^3_{\hat{c}^3}$ and the descriptor $\phi_h(S)$ is of the weighted average type, we could identify the composition ratio $(\hat{c}^1, \hat{c}^2, \hat{c}^3)$ that approximately satisfies the following condition:

$$\mathcal{C}_h = \left\{ (\hat{c}^1, \hat{c}^2, \hat{c}^3) \middle| \sum_{i=1}^{3} \hat{c}^i \eta(S^i) = m_h, \sum_{i=1}^{3} \hat{c}^i = 1, \hat{c}^i \geq 0 \ (\forall i) \right\},$$

where $\hat{c}^i$ denotes the normalized fraction and $\eta(S^i)$ is the feature value of element S^i. Without any loss of generality, $\mathcal{C}_h$ can be defined for any system or other descriptor type such as the weighted variance. Here, we focused on the weighted average descriptors of the van der Waals radius ("ave:dw_radius_uff"), Ghosh's scale of electronegativity ("ave:en_ghosh"), first ionization energy ("ave:first_ion_en"), number of filled p valence orbitals ("ave:num_p_valence"), and energy per atom in the $T = 0$K ground state calculated by density functional theory ("ave:gs_energy") among the highly relevant descriptors listed in Table 3. Then, we overwrote each $\mathcal{C}_h$ on the predicted phase diagrams for the 30 alloy systems. Figure 6 illustrates eight selected phase diagrams (see also Figure S2 in the Supplementary Note for the results of all 30 systems). In almost all systems, the straight lines $\mathcal{C}_h$ conditioned by the five relevant descriptors passed through the predicted QC and AC phase regions. Note that each $\mathcal{C}_h$ is one of the necessary conditions for the formation of QC and AC phases. The intersection of these conditions defines a set of empirical equations for determining the compositional ratio that forms a quasicrystal or approximant.

In this way, the implicit rules extracted by the machine learning algorithm can be encoded in a simple mathematical form interpretable by humans. By accumulating such empirical rules, performing verifications, and pursuing theoretical explanations, we can gain new scientific knowledge. It is important to note that the empirical equations described here are subject to various restrictions in terms of their applicable domains. Specifically, they may be local rules obtained from the input–output of the trained model for ternary alloys of Al-TM-TM and thus would not be generally applicable to other systems. There must be many other implicit rules to discover from the trained model, and thus it is important to exhaustively extract these implicitly encoded rules and clarify their range of application at the same time.

Table 3: The 20 most relevant descriptors in the classification task for the 30 Al-TM-TM alloy systems. The first column shows the descriptor name (upper) and ID (lower) in XenonPy. The prefixes "ave" and "var" in the descriptor ID represent weighted average and weighted variance types, respectively. The last four columns show the within-class means of the QC, AC, others, and QC/AC-merged groups. The within-class variances (converted to standard deviations) are reported in parentheses.

Descriptor information	MIC	QC	AC	others	QC/AC
ave:vdw_radius_uff van der Waals radius from the UFF (pm)	0.43	409.05 (3.37)	406.49 (6.81)	382.30 (40.66)	406.59 (6.73)
ave:en_ghosh Ghosh's scale of electronegativity	0.42	0.15 (0.00)	0.15 (0.01)	0.16 (0.02)	0.15 (0.01)
ave:first_ion_en First ionization energy (eV)	0.41	6.49 (0.09)	6.53 (0.17)	6.84 (0.58)	6.53 (0.17)
ave:mendeleev_number Mendeleev's number	0.41	75.94 (0.36)	75.86 (1.47)	73.32 (4.33)	75.87 (1.45)
ave:specific_heat Specific heat at 20 °C (J/(g mol))	0.40	0.74 (0.02)	0.73 (0.04)	0.66 (0.15)	0.73 (0.04)
ave:num_p_valence Number of filled p valence orbitals	0.40	0.71 (0.06)	0.73 (0.05)	0.57 (0.25)	0.73 (0.05)
ave:num_p_unfilled Number of unfilled p valence orbitals	0.40	3.53 (0.30)	3.63 (0.24)	2.85 (1.24)	3.63 (0.24)
ave:heat_capacity_mass Specific heat capacity at STP (J/mol-K)	0.40	0.74 (0.02)	0.73 (0.04)	0.66 (0.15)	0.73 (0.04)
ave:covalent_radius_cordero Covalent radius by Cerdero et al. (pm)	0.39	126.06 (1.19)	126.38 (2.13)	129.51 (6.31)	126.37 (2.10)
ave:vdw_radius van der Waals radius (pm)	0.37	189.51 (0.55)	190.67 (1.81)	193.80 (6.33)	190.63 (1.79)
ave:gs_energy Ground state energy at T=0K (eV/atom)	0.37	-4.57 (0.18)	-4.69 (0.28)	-5.19 (1.12)	-4.68 (0.28)
ave:thermal_conductivity Thermal conductivity at 25°C (W/(m K))	0.36	221.35 (21.74)	201.23 (13.96)	170.72 (60.68)	201.99 (14.83)
ave:covalent_radius_slater Covalent radius by Slater (pm)	0.35	127.92 (1.19)	128.08 (0.93)	130.40 (3.38)	128.08 (0.94)
ave:period Period in periodic table	0.35	3.40 (0.06)	3.52 (0.19)	3.73 (0.55)	3.52 (0.19)
var:num_p_valence Number of filled p valence orbitals (pm)	0.34	0.20 (0.02)	0.20 (0.02)	0.18 (0.07)	0.20 (0.02)
ave:num_d_valence Number of filled d valence orbitals (pm)	0.34	2.30 (0.60)	2.08 (0.52)	3.15 (1.90)	2.09 (0.52)
ave:heat_capacity_molar Molar heat capacity at STP (J/mol-K)	0.34	24.44 (0.10)	24.47 (0.14)	24.81 (0.58)	24.47 (0.14)
ave:density Density at 295K (g/cm^3)	0.34	4.94 (0.32)	5.55 (1.24)	6.70 (3.35)	5.53 (1.23)
var:num_p_unfilled Number of unfilled p valence orbitals	0.34	5.09 (0.60)	4.91 (0.48)	4.60 (1.77)	4.91 (0.49)
ave:hhi_p Herfindahl-Hirschman Index (HHI) production values	0.33	1810.99 (242.60)	2106.51 (274.67)	2196.88 (706.35)	2095.30 (279.22)

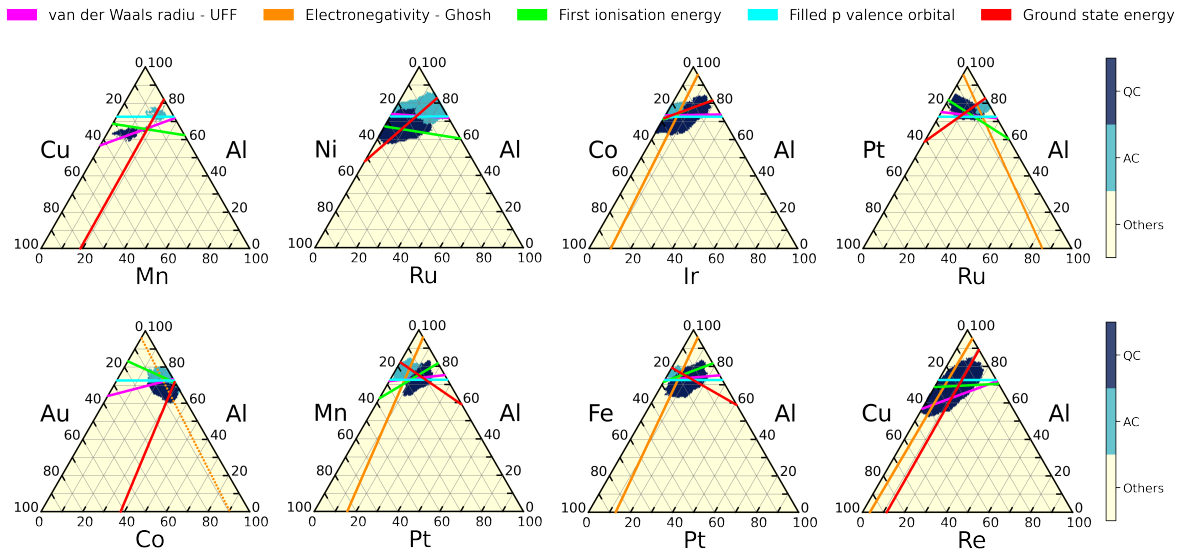

Figure 6: Five rules for the formation of QC/AC phases proposed by machine learning. The rules are represented by straight lines on the predicted phase diagrams of the eight systems. Each line represents a condition set $\mathcal{C}_h$ describing the weighted average of the van der Waals radius, electronegativity, first ionization energy, number of filled p valence orbitals, or energy per atom in the $T = 0$K ground state that is imposed on the compositional formula.

Discussion

This study demonstrated the predictive power of machine learning for the identification of candidate compositions to form quasicrystalline and approximant alloys. The problem was formulated rather simply as a supervised learning task of classifying any given composition into one of three kinds of material structures: quasicrystals, approximants, and others representing types other than the first two. Although supervised learning was conducted with a conventional random forest classifier, the model trained only on a list of known compositions reached a high prediction accuracy. In a binary classification task of predicting a combined quasicrystal/approximant class versus others, the precision and recall reached 0.997 and 0.999, respectively. In addition, it was confirmed that the model can discriminate between quasicrystals and approximants, although the accuracy is slightly lower. If this approach can be used to narrow candidate compositions for forming quasicrystals and approximants, the efficiency of related materials searches would be greatly improved.

The predictability of machine learning for this task has been proven. However, before putting the approach into practice, some remaining questions need to be answered. The first is why the machine learning models can predict the compositions of quasicrystals. In the present study, we evaluated the relevance of the descriptors based on the MIC metric and narrowed the total to five descriptors that are strongly involved in the model decision making process. According to the identified descriptors, we derived five empirical equations with high interpretability that are presumed to be necessary conditions for the formation and hence the stability of quasicrystals and approximants. Importantly, these newly identified conditions will lead to the long-sought and heretofore unclear guiding principles for the synthesis of new quasicrystals, thereby opening the door to a deeper understanding of quasicrystal stability as a central issue in condensed matter physics. Many other implicit rules are still embedded in the learned model. By identifying the comprehensive set of rules encoded in the black-box machine learning model, we will piece together the puzzle and record statements as rules of thumb for materials science.

The other remaining question concerns the applicable domains of these machine learning models. Most of the quasicrystals found thus far are binary or ternary systems. In fact, there are only 12 quasicrystals of quaternions or more in our training dataset. It is expected that stable quasicrystals will be more likely to form from systems consisting of a greater number of elements since, for instance, the number of ternary quasicrystals is much larger than the number of binary quasicrystals. On the other hand, predictions based on data science technologies are interpolative by nature, and thus it is now of particular interest to determine to what extent models trained primarily from binary and ternary systems can be generalized for multidimensional systems where less or no data are available.

With this study, we have taken the first step in the practical application of data science toward the accelerated discovery of new quasicrystals. However, there are still some technical improvements to be made. To facilitate

subsequent research, we have published all datasets that were used for machine learning and benchmarking. With these datasets, all results shown in this paper can be reproduced on our platform, XenonPy. This is expected to promote comprehensive experimental validation in the quasicrystal research community.

Methods

Data preparation

The list of 80 quasicrystals and 78 approximants was compiled from the Crystallography of Quasicrystals handbook. In addition, 10,090 compositions of ordinary crystals were extracted from the Materials Project database and laboratory data on failed quasicrystal syntheses. One of the difficulties in model building arose from the bias in the number of samples in different classes: 80 and 78 compositions for quasicrystals and approximants, respectively, as opposed to 126,335 crystals from the Materials Project database (V2020.08.20). Therefore, to manage the highly unbalanced class labels, we down-sampled the crystal data by randomly extracting 10,000 instances from the overall data taken from the Materials Project database.

To evaluate the validity of the predicted phase diagrams, we gathered 30 experimental phase diagrams of Al-TM-TM alloy systems from 25 papers. To facilitate the collection, an in-house software was developed to accelerate data extraction from published phase diagram images. We quantified the difference and overlap between the extracted phase regions and predicted quasicrystalline and approximant phase regions to evaluate the true positive and false positive rates as detailed in the Results section.

Compositional pattern of datasets

Figure 1 (b) shows a low-dimensional representation of the compositional distribution of our data belonging to the three classes, which was used to determine the between-class difference and overlap. The compiled list of quasicrystals and approximants consisted of 26 binary, 120 ternary, and 12 quaternary systems spanning 50 different elements. On the other hand, the ordinary crystal dataset consisted of unitary to octogenarian systems with constituents spanning a broader range of elements. To more clearly visualize the difference and overlap in the class-specific distributions, only binary to quaternary crystals are shown on the plot. Furthermore, crystals containing elements other than the constituents of the quasicrystals and approximants are excluded. We translated each composition into a 50-dimensional binary vector with each entry encoding the presence or absence of an element as one or zero, respectively. The feature vectors of 19,191 compositions were projected onto a two-dimensional subspace using a dimensionality reduction technique called UMAP [24]. There is no significant bias in the distribution of the three classes at the level of their constituent elements, implying that no particular combination of elements is favorable for the formation of quasicrystals. The visualized data pattern also suggests that previous studies on quasicrystals have explored a wide range of compositional spaces without bias toward any particular compositional combination.

Random forest classifier

A random forest classifier was built on an ensemble of decision tree models. The overall dataset was randomly divided into training and test sets as described in the Results section. We performed cross-validation in the training dataset and selected the hyperparameters that minimized the prediction error. The hyperparameters and search candidates are summarized in Table 4. The number of combinations of search candidates was 96. As mentioned in the Results section, the training dataset consisted of 66 quasicrystals and 60 approximant crystals. This dataset contained 69 unique ternary systems. In the cross-validation, the compositional data belonging to each ternary system were lumped together, and the training and validation datasets were divided based on the ternary systems; i.e., one of the 69 systems was used as the validation set, and all the remaining data, including the data outside the 69 systems, were used for training. To quantify the prediction uncertainty, we also trained models from 100 randomly selected datasets with the selected hyperparameters. Using these models, we calculated the mean and standard deviation of the performance metrics with respect to the test dataset. The learning algorithm implemented in scikit-learn [55] v0.23.1 (https://github.com/scikit-learn/scikit-learn/releases/tag/0.23.1) was employed to train the models.

Data availability

The authors confirm that the data supporting the findings of this study are available in the Supplementary Information, including the digital data of the quasicrystals and approximant crystals. The list of ordinary crystals is accessible at the Materials Project database. Digital data for the 30 experimental phase diagrams

Table 4: List of hyperparameters and their search candidates (grid points and module options of scikit-learn) used for cross-validation. The selected combination of hyperparameters is shown in bold.

Hyperparameter	Search candidate
Number of trees (n_estimators)	100, **200**, 300
Maximum depth of trees (max_depth)	10, 15, 20, **25**
Number of features in each tree (max_features)	sqrt, **$\log_2$**
Bootstrap sampling in the bagging	**False**, True
Classification loss	**entropy**, gini

are available on request from the corresponding authors. With these datasets, all results of this study can be reproduced with the XenonPy software.

References

[1] Shechtman, D., Blech, I., Gratias, D. & Cahn, J. W. Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* **53**, 1951–1953 (1984). doi:10.1103/PhysRevLett.53.1951.

[2] Tsai, A. P., Inoue, A. & Masumoto, T. A stable quasicrystal in Al-Cu-Fe system. *Jpn. J. Appl. Phys.* **26**, L1505–L1507 (1987). doi:10.1143/JJAP.26.L1505.

[3] Tsai, A.-P., Inoue, A. & Masumoto, T. Stable decagonal Al-Co-Ni and Al-Co-Cu quasicrystals. *Mater. Trans. JIM* **30**, 463–473 (1989). doi:10.2320/matertrans1989.30.463.

[4] Tsai, A. P., Inoue, A., Yokoyama, Y. & Masumoto, T. Stable icosahedral Al-Pd-Mn and Al-Pd-Re alloys. *Mater. Trans. JIM* **31**, 98–103 (1990). doi:10.2320/matertrans1989.31.98.

[5] Takakura, H., Gómez, C. P., Yamamoto, A., De Boissieu, M. & Tsai, A. P. Atomic structure of the binary icosahedral Yb–Cd quasicrystal. *Nat. Mater.* **6**, 58–63 (2007). doi:10.1038/nmat1799.

[6] Yamada, T., Takakura, H., de Boissieu, M. & Tsai, A.-P. Atomic structures of ternary Yb–Cd–Mg icosahedral quasicrystals and a 1/1 approximant. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **73**, 1125–1141 (2017). doi:10.1107/S2052520617013270.

[7] Kimura, K. *et al.* Electronic properities of the single-grained icosahedral phase of Al–Li–Cu. *J. Phys. Soc. Japan* **58**, 2472–2481 (1989). doi:10.1143/JPSJ.58.2472.

[8] Mizutani, U. *et al.* Electron transport properties of thermodynamically stable Al-Cu-Ru icosahedral quasicrystals. *J. Phys. Condens. Matter* **2**, 6169–6178 (1990). doi:10.1088/0953-8984/2/28/007.

[9] Akiyama, H., Honda, Y., Hashimoto, T., Edagawa, K. & Takeuchi, S. Toward insulating quasicrystalline alloy in Al-Pd-Re icosahedral phase. *Jpn. J. Appl. Phys.* **32**, L1003–L1004 (1993). doi:10.1143/JJAP.32.L1003.

[10] Watanuki, T. *et al.* Intermediate-valence icosahedral Au-Al-Yb quasicrystal. *Phys. Rev. B* **86**, 094201 (2012). doi:10.1103/PhysRevB.86.094201.

[11] Deguchi, K. *et al.* Quantum critical state in a magnetic quasicrystal. *Nat. Mater.* **11**, 1013–1016 (2012). doi:10.1038/nmat3432.

[12] Kamiya, K. *et al.* Discovery of superconductivity in quasicrystal. *Nat. Commun.* **9**, 154 (2018). doi:10.1038/s41467-017-02667-x.

[13] Hume-Rothery, W. Research on the nature, properties and conditions of formation of intermetallic compounds, with special reference to certain compounds of Tin. *J. Inst. Metals* **35**, 295–299 (1926).

[14] Tsai, A. P. Icosahedral clusters, icosaheral order and stability of quasicrystals - a view of metallurgy. *Sci. Technol. Adv. Mater.* **9**, 013008 (2008). doi:10.1088/1468-6996/9/1/013008.

[15] Nakayama, K., Mizutani, A. & Koyama, Y. Crystallographic features and state stability of the decagonal quasicrystal in the Al–Co–Cu alloy system. *J. Phys. Soc. Japan* **85**, 114602 (2016). doi:10.7566/JPSJ.85.114602.

[16] Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016). doi:10.1038/nmat4717.

[17] Hansen, E. C. *et al.* New ligands for nickel catalysis from diverse pharmaceutical heterocycle libraries. *Nat. Chem.* **8**, 1126–1130 (2016). doi:10.1038/nchem.2587.

[18] Sumita, M., Yang, X., Ishihara, S., Tamura, R. & Tsuda, K. Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies. *ACS Cent. Sci.* **4**, 1126–1133 (2018). doi:10.1021/acscentsci.8b00213.

[19] Wu, S. *et al.* Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **5**, 66 (2019). doi:10.1038/s41524-019-0203-2.

[20] Oliynyk, A. O. *et al.* High-throughput machine-learning-driven synthesis of full-Heusler Compounds. *Chem. Mater.* **28**, 7324–7331 (2016). doi:10.1021/acs.chemmater.6b02724.

[21] Matsumoto, R. *et al.* Two pressure-induced superconducting transitions in SnBi2Se4 explored by data-driven materials search: new approach to developing novel functional materials including thermoelectric and superconducting materials. *Appl. Phys. Express* **11**, 093101 (2018). doi:10.7567/APEX.11.093101. 1806.09284.

[22] Seko, A. *et al.* Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and Bayesian optimization. *Phys. Rev. Lett.* **115**, 205901 (2015). doi:10.1103/PhysRevLett.115.205901.

[23] Carrete, J., Li, W., Mingo, N., Wang, S. & Curtarolo, S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys. Rev. X* **4**, 011019 (2014). doi:10.1103/PhysRevX.4.011019. 1401.2439.

[24] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018). doi:10.21105/joss.00861.

[25] Steurer, W. & Deloudi, S. *Crystallography of Quasicrystals*, vol. 126 of *Springer Series in Materials Science* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009). doi:10.1007/978-3-642-01899-2.

[26] Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013). doi:10.1063/1.4812323.

[27] Xenonpy platform. https://github.com/yoshida-lab/XenonPy. Accessed: 2021-1-20.

[28] Mi, S., Grushko, B., Dong, C. & Urban, K. A study of the ternary phase diagrams of Al–Co with Cu, Ag and Au. *J. Alloys Compd.* **354**, 148–152 (2003). doi:10.1016/S0925-8388(02)01356-7.

[29] Grushko, B. & Velikanova, T. Formation of quasiperiodic and related periodic intermetallics in alloy systems of aluminum with transition metals. *Comput. Coupling Phase Diagrams Thermochem.* **31**, 217–232 (2007). doi:10.1016/j.calphad.2006.12.002.

[30] Pavlyuchkov, D., Grushko, B. & Velikanova, T. An investigation of the Al–Pd–Ir phase diagram between 50 and 100at.% Al. *J. Alloys Compd.* **453**, 191–196 (2008). doi:10.1016/j.jallcom.2006.11.163.

[31] Grushko, B., Kowalski, W., Pavlyuchkov, D., Mi, S. & Surowiec, M. Al-rich region of the Al–Ni–Cr alloy system below 900°C. *J. Alloys Compd.* **485**, 132–138 (2009). doi:10.1016/j.jallcom.2009.05.093.

[32] Grushko, B., Kowalski, W. & Balanetskyy, S. Phase equilibria in the Al-rich region of the Al–Ni–Re alloy system. *J. Alloys Compd.* **479**, L59–L61 (2009). doi:10.1016/j.jallcom.2009.01.084.

[33] Grushko, B., Kowalski, W. & Surowiec, M. On the constitution of the Al–Co–Fe alloy system. *J. Alloys Compd.* **491**, L5–L7 (2010). doi:10.1016/j.jallcom.2009.10.156.

[34] Kowalski, W., Grushko, B., Pavlyuchkov, D. & Surowiec, M. A contribution to the Al-Pd-Cr phase diagram. *J. Alloys Compd.* **496**, 129–134 (2010). doi:10.1016/j.jallcom.2010.02.033.

[35] Kapush, D., Velikanova, T. & Grushko, B. An investigation of the Al-rich region of the Al–Ni–Ir phase diagram. *J. Alloys Compd.* **497**, 105–109 (2010). doi:10.1016/j.jallcom.2010.03.056.

[36] Grushko, B., Kapush, D., Velikanova, T. Y., Samuha, S. & Meshi, L. An investigation of the Al-Rh-Ru phase diagram above 50 at.% Al. *J. Alloys Compd.* **509**, 8018–8021 (2011). doi:10.1016/j.jallcom.2011.05.074.

[37] Grushko, B. & Mi, S. A study of the Al-rich region of the Al–Cu–Mo alloy system. *J. Alloys Compd.* **509**, L30–L33 (2011). doi:10.1016/j.jallcom.2010.10.001.

[38] Grushko, B., Kapush, D., Konoval, V. & Shemet, V. A study of the Al-Ni-Pt alloy system. Phase equilibria at 1100 and 1300°C. *Powder Metall. Met. Ceram.* **50**, 462–470 (2011). doi:10.1007/s11106-011-9350-9.

[39] Balanetskyy, S., Meisterernst, G., Grushko, B. & Feuerbacher, M. The Al-rich region of the Al-Mn-Ni alloy system. Part II. Phase equilibria at 620-1000°C. *J. Alloys Compd.* **509**, 3795–3805 (2011). doi:10.1016/j.jallcom.2010.10.114.

[40] Zaikina, O. V., Khorujaya, V. G., Pavlyuchkov, D., Grushko, B. & Velikanova, T. Y. Investigation of the Al-Ti-Pd alloy system at 930 and 1100°C. *J. Alloys Compd.* **509**, 43–51 (2011). doi:10.1016/j.jallcom.2010.08.144.

[41] Grushko, B., Kapush, D. & Meshi, L. A study of the Al-rich part of the Al-Ni-Pt alloy system. *J. Alloys Compd.* **514**, 60–63 (2012). doi:10.1016/j.jallcom.2011.10.076.

[42] Grushko, B., Kapush, D., Samuha, S. & Meshi, L. A study of the Al-Pd-Pt alloy system. *J. Alloys Compd.* **600**, 125–129 (2014). doi:10.1016/j.jallcom.2014.02.109.

[43] Grushko, B. A study of phase equilibria in the Al-Pt-Rh alloy system. *J. Alloys Compd.* **636**, 329–334 (2015). doi:10.1016/j.jallcom.2015.02.116.

[44] Kapush, D., Samuha, S., Meshi, L., Velikanova, T. Y. & Grushko, B. Formation of complex intermetallics in the Al-rich part of Al-Pt-Ru. *J. Phase Equilibria Diffus.* **36**, 327–332 (2015). doi:10.1007/s11669-015-0385-3.

[45] Grushko, B., Pavlyuchkov, D., Mi, S. & Balanetskyy, S. Ternary phases forming adjacent to Al 3 Mn Al 4 Mn in Al Mn TM (TM = Fe, Co, Ni, Cu, Zn, Pd). *J. Alloys Compd.* **677**, 148–162 (2016). doi:10.1016/j.jallcom.2016.03.220.

[46] Grushko, B. & Mi, S. B. Al-rich region of Al–Cu–Mn. *J. Alloys Compd.* **688**, 957–963 (2016). doi:10.1016/j.jallcom.2016.07.075.

[47] Samuha, S., Grushko, B. & Meshi, L. Refinement of the Al-rich part of the Al–Cu–Re phase diagram and atomic model of the ternary $Al_{6.2}Cu_2Re$ phase. *J. Alloys Compd.* **670**, 18–24 (2016). doi:10.1016/j.jallcom.2016.02.070.

[48] Grushko, B. A contribution to the Al–Cu–Cr phase diagram. *J. Alloys Compd.* **729**, 426–437 (2017). doi:10.1016/j.jallcom.2017.09.116.

[49] Grushko, B., Kowalski, W. & Mi, S. A study of the Al–Co–Cr alloy system. *J. Alloys Compd.* **739**, 280–289 (2018). doi:10.1016/j.jallcom.2017.12.226.

[50] Grushko, B. A contribution to the ternary phase diagrams of Al with Co, Rh and Ir. *J. Alloys Compd.* **772**, 399–408 (2019). doi:10.1016/j.jallcom.2018.09.066.

[51] Grushko, B. A study of the Al–Mn–Pt alloy system. *J. Alloys Compd.* **792**, 1223–1229 (2019). doi:10.1016/j.jallcom.2019.04.130.

[52] Grushko, B. A study of the Al–Fe–Pt alloy system. *J. Alloys Compd.* **829**, 154444 (2020). doi:10.1016/j.jallcom.2020.154444.

[53] Kitahara, K. & Kimura, K. Local cluster networks and the number of valence states in aluminium–transition metal face-centred icosahedral quasicrystals. *Zeitschrift für Krist. - Cryst. Mater.* **232**, 507–513 (2017). doi:10.1515/zkri-2016-2035.

[54] Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011). doi:10.1126/science.1205438. https://science.sciencemag.org/content/334/6062/1518.full.pdf.

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

Acknowledgements

This work was supported in part by a MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number 19H50820). Ryo Yoshida acknowledges the financial support from a Grant-in-Aid for Scientific Research (A) 19H01132 from the Japan Society for the Promotion of Science (JSPS) and JST CREST Grant Number JPMJCR19I3.

Author contributions

Ryo Yoshida and Kaoru Kimura designed the conceptual idea and proof outline. Chang Liu and Ryo Yoshida wrote the manuscript and carried out the data analysis. Erina Fujita, Yukari Katsura, Yuki Inada, Asuka Ishikawa, Ryuji Tamura, and Kaoru Kimura worked out the collection and curation of the dataset. All authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information

- **Supplementary Data** List of quasicrystals and approximants that were used in the supervised learning. Each column denotes the chemical composition, class label indicating quasicrystal or approximant (QC and AC), and structure type of the QC or AC.
- **Supplementary Note** Supplementary Figure S1-S2 and Supplementary Table S1-S3.

Figures

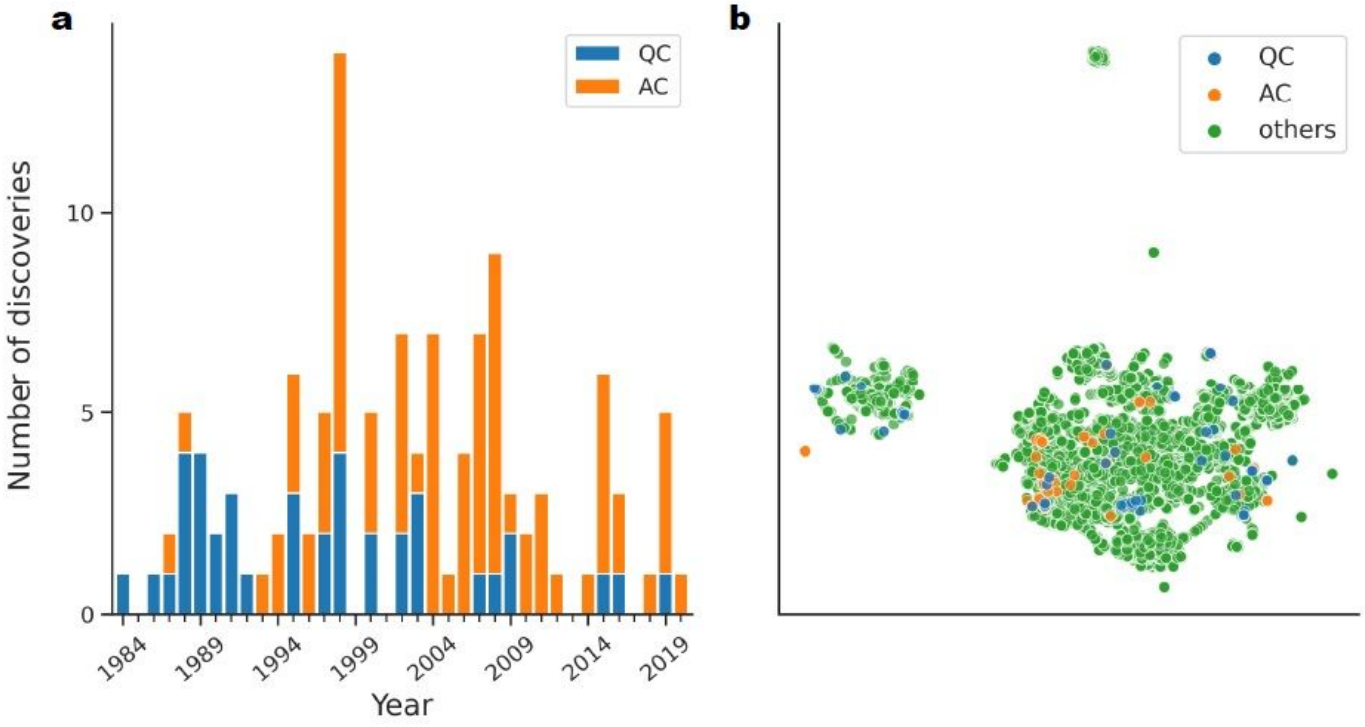

Figure 1

Quasicrystals (QC) and approximants (AC) that have been discovered so far. a. Annual trend in the discovery of new quasicrystals (blue) and approximant crystals (orange) in aluminum alloys. b. Distribution of the compositional dataset that was visualized onto a two-dimensional space obtained by the UMAP algorithm [24] (see the Methods section). Quasicrystals, approximants, and ordinary crystals are color-coded by blue, orange, and green, respectively.

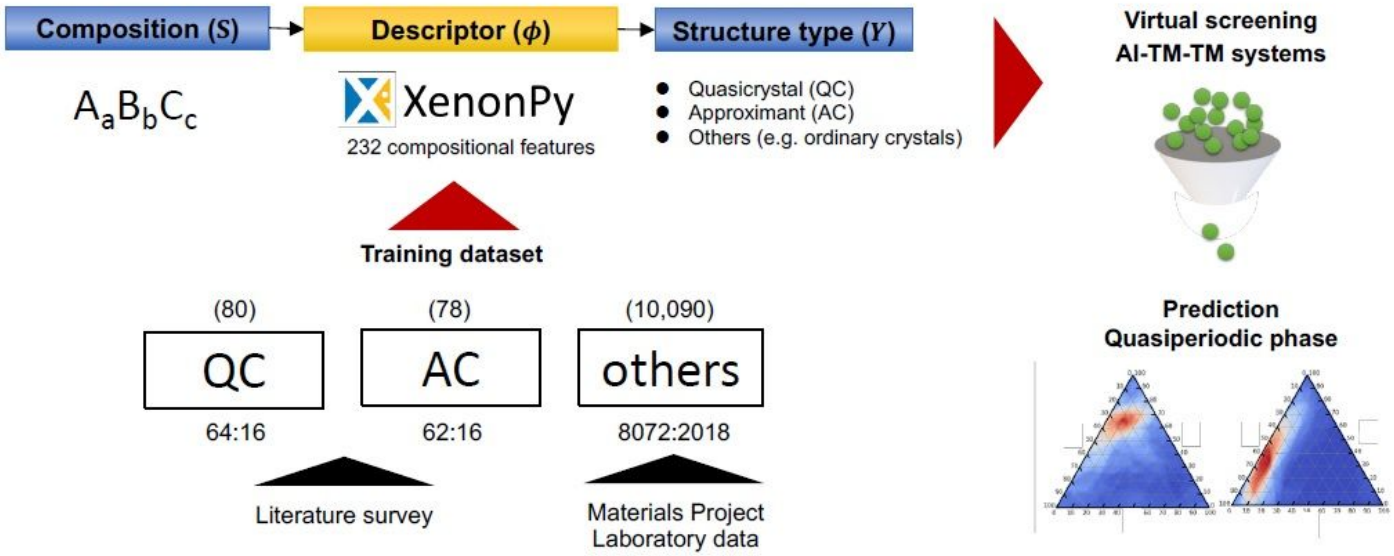

Figure 2

Machine learning workflow. The compositional features were encoded into a 232-dimensional descriptor vector, and a prediction model was created using a random forest classifier. The trained model predicts the class label of a given chemical composition as quasicrystal (QC), approximant (AC), or others. Model training and testing were performed on the compositional features of 80 known quasicrystals, 78 approximants, and 10,090 ordinary crystals. Finally, we performed HTS across all Al-TM-TM (TM: transition metal) alloys to generate their predicted phase diagrams. The results were compared with experimental phase diagrams obtained from the literature.

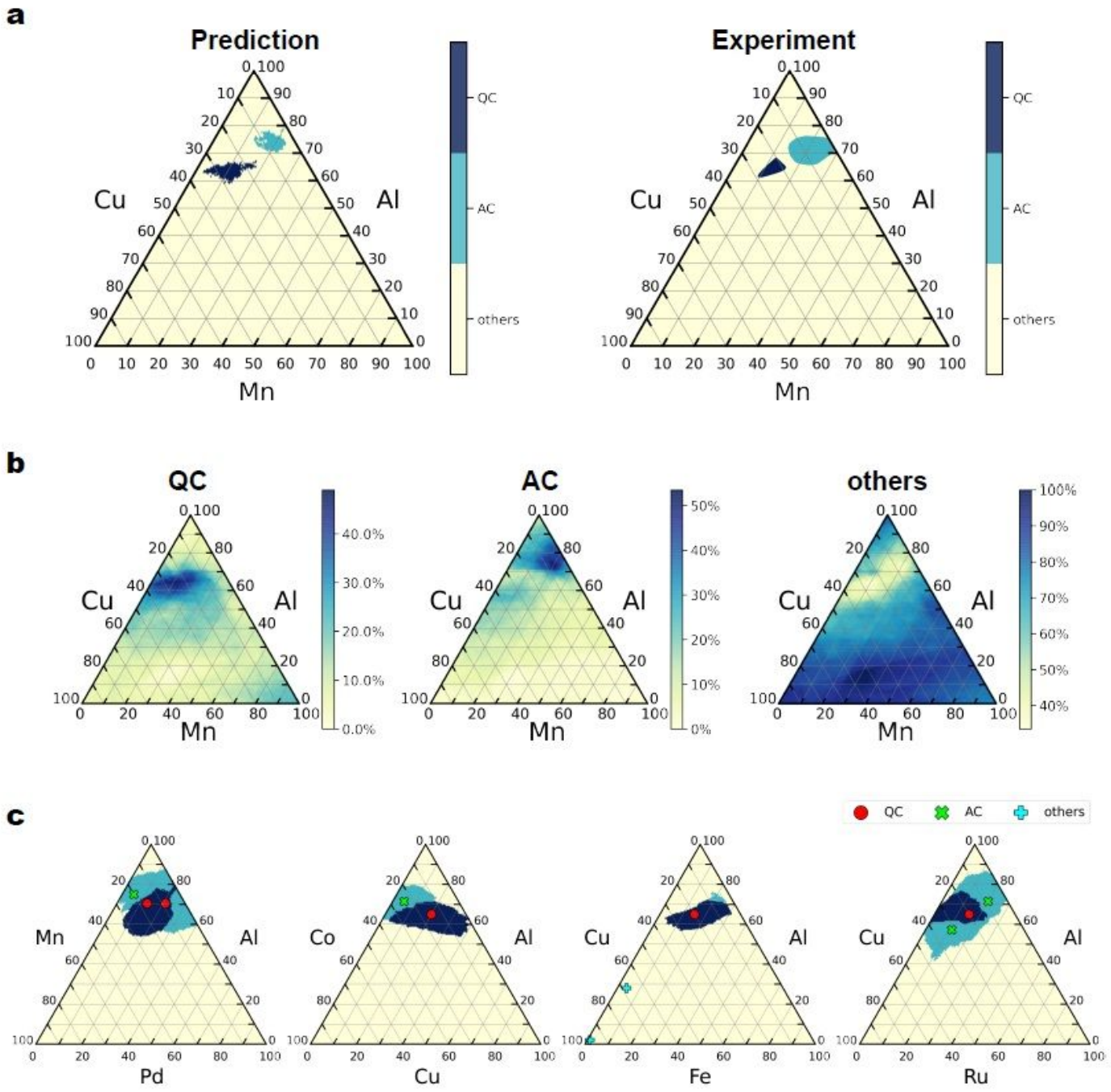

Figure 3

Phase prediction of the Al-Mn-Cu system. a. Predicted phase diagram (left panel) and experimental phase diagram (right panel) of the Al-Mn-Cu system. The three colors denote the quasicrystalline phase (QC), approximant phase (AC), and others. Despite the lack of training instances for Al-Mn-Cu, the model successfully predicts the unseen quasicrystalline and approximant crystalline phases. b. Heatmap display of the predicted class probability of QC, AC, and others for the Al-Mn-Cu system. c. In order to observe the training instances relevant to the model decision making, we examined the distribution of training instances in the four ternary systems closest to Al-Mn-Cu.

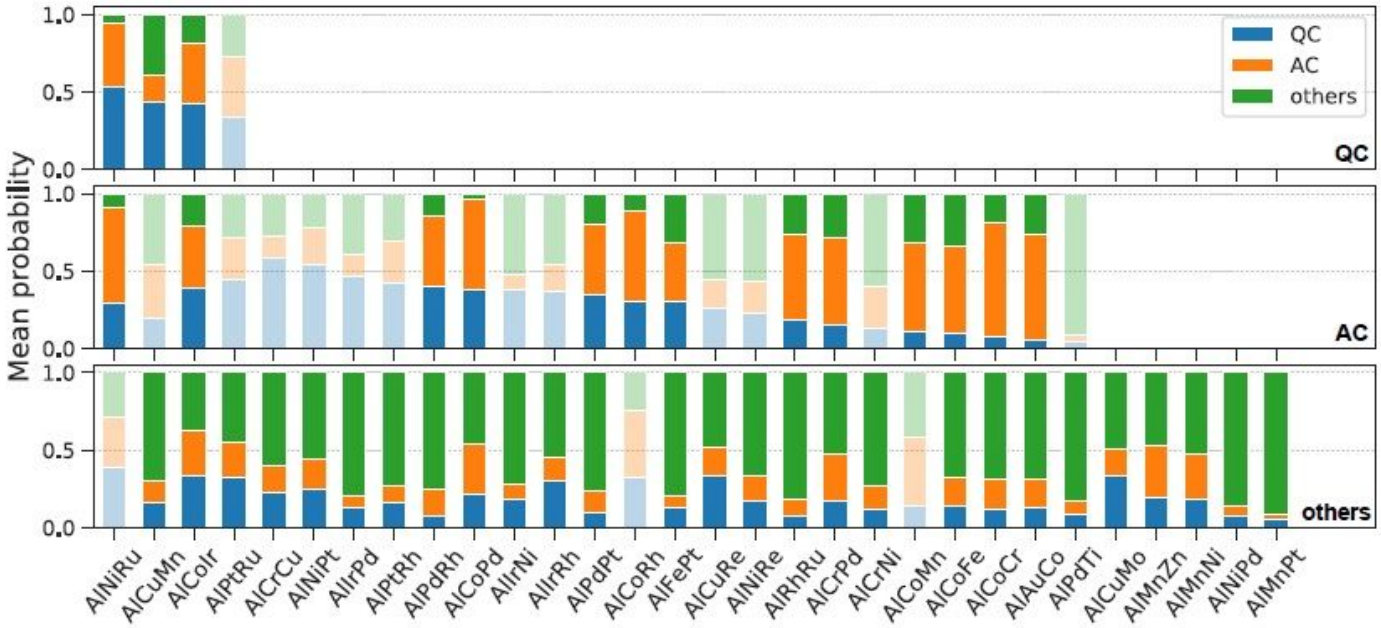

Figure 4

Prediction performance for the 30 different Al-TM-TM systems. The mean class probability was calculated in each of the experimental phase regions (top: QC, middle: AC, and bottom: others) using our trained random forest classifier. The bar plots shown in a transparent color represent phases where class label prediction based on maximum probability failed.

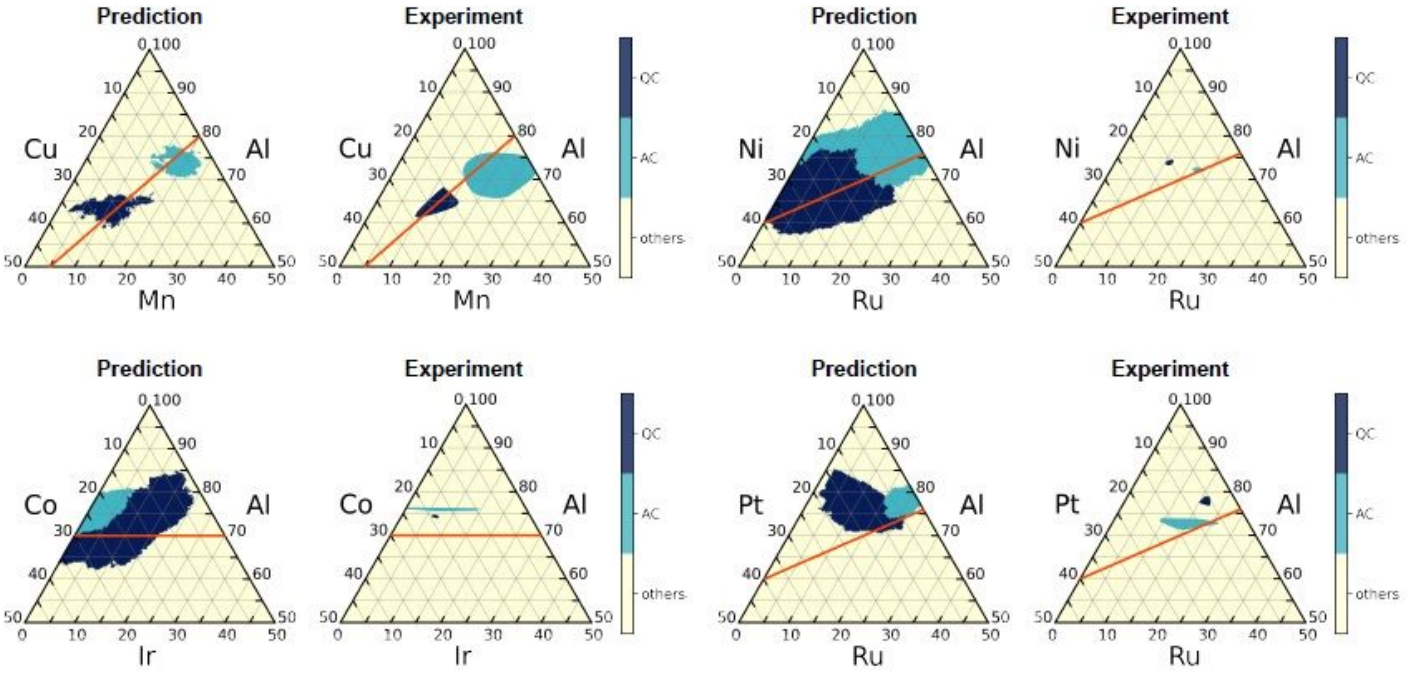

Figure 5

Predicted and experimental phase diagrams of four ternary alloy systems. The orange lines indicate the Hume-Rothery rule of valence electron concentration with e/a = 1:8.

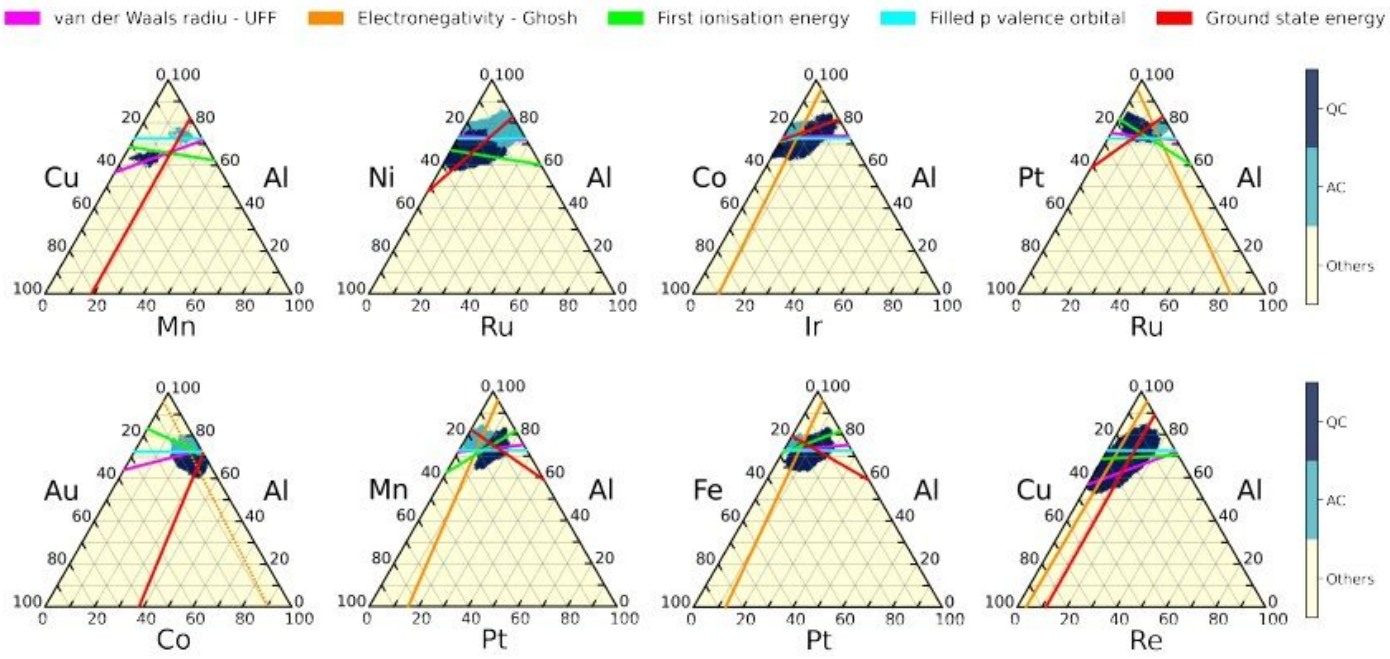

Figure 6

Five rules for the formation of QC/AC phases proposed by machine learning. The rules are represented by straight lines on the predicted phase diagrams of the eight systems. Each line represents a condition set Ch describing the weighted average of the van der Waals radius, electronegativity, first onization energy,

number of filled p valence orbitals, or energy per atom in the T = 0K ground state that is imposed on the compositional formula.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- quasicrystalandapproximant.csv
- MLonquasicrystalsnaturecommunicationsSI.pdf