

Machine Learning to Predict Quasicrystals from Chemical Compositions

Chang Liu, Erina Fujita, Yukari Katsura, Yuki Inada, Asuka Ishikawa, Ryuji Tamura, Kaoru Kimura,* and Ryo Yoshida*

Quasicrystals have emerged as the third class of solid-state materials, distinguished from periodic crystals and amorphous solids, which have long-range order without periodicity exhibiting rotational symmetries that are disallowed for periodic crystals in most cases. To date, more than one hundred stable quasicrystals have been reported, leading to the discovery of many new and exciting phenomena. However, the pace of the discovery of new quasicrystals has lowered in recent years, largely owing to the lack of clear guiding principles for the synthesis of new quasicrystals. Here, it is shown that the discovery of new quasicrystals can be accelerated with a simple machine-learning workflow. With a list of the chemical compositions of known stable quasicrystals, approximant crystals, and ordinary crystals, a prediction model is trained to solve the three-class classification task and its predictability compared to the observed phase diagrams of ternary aluminum systems is evaluated. The validation experiments strongly support the superior predictive power of machine learning, with the overall prediction accuracy of the phase prediction task reaching ≈ 0.728 . Furthermore, analyzing the input–output relationships black-boxed into the model, nontrivial empirical equations interpretable by humans that describe conditions necessary for stable quasicrystal formation are identified.


1. Introduction

This study demonstrates the potential of machine learning to predict stable 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 Shechtman 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 stable 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 1a** shows the annual trend of new stable quasicrystals found in aluminum alloy systems. From 1986 to 1999, new stable 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 stable quasicrystals. In terms of the stability mechanism of quasicrystals, the Hume-Rothery rules,^[13] that is, itinerant valence electron concentration, e/a , and atomic size factor, have been considered.^[14,15] However, these are only necessary conditions and are insufficient on their own. Thus, we aimed to accelerate the discovery of new stable quasicrystals by introducing machine learning to the field.

Dr. C. Liu, Prof. R. Yoshida
The Institute of Statistical Mathematics
Research Organization of Information and Systems
Tachikawa 190-8562, Japan
E-mail: yoshidar@ism.ac.jp

E. Fujita, Dr. Y. Katsura, Y. Inada, K. Kimura
Department of Advanced Materials Science
The University of Tokyo
Kashiwa 277-8561, Japan
E-mail: bkimura@phys.mm.t.u-tokyo.ac.jp

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adma.202102507>.

© 2021 The Authors. Advanced Materials published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/adma.202102507

A. Ishikawa, Prof. R. Tamura
Department of Materials Science and Technology
Tokyo University of Science
Tokyo 125-8585, Japan

Prof. R. Yoshida
Research and Service Division of Materials Data and Integrated System
National Institute for Materials Science
Tsukuba 305-0047, Japan

Prof. R. Yoshida
Department of Statistical Science
The Graduate University for Advanced Studies
Tachikawa 190-8562, Japan

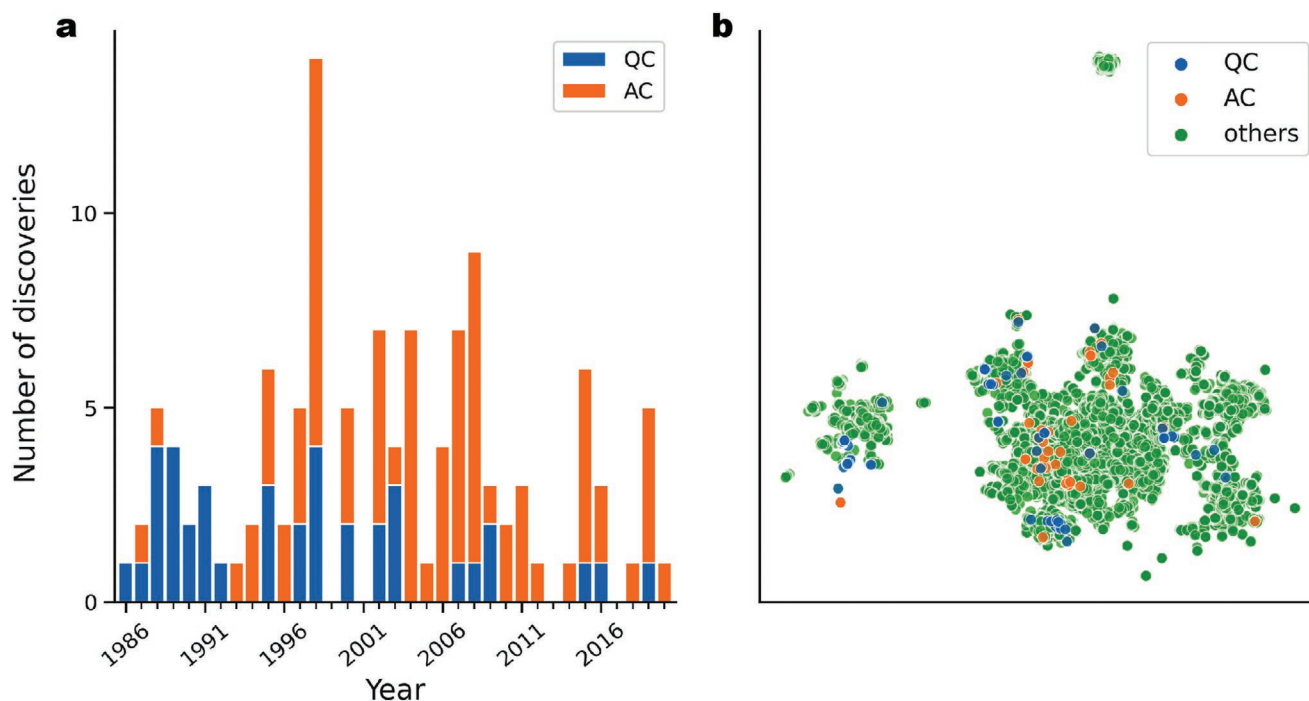


Figure 1. Stable quasicrystals (QC) and approximants (AC) that have been discovered so far. a) Annual trend in the discovery of new thermodynamically stable quasicrystals (blue) and approximant crystals (orange) in aluminum alloys. b) Distribution of the compositional dataset that was visualized onto a 2D space obtained by the UMAP algorithm^[30] (see the Experimental Section). Stable quasicrystals, approximants, and ordinary crystals are color-coded by blue, orange, and green, respectively.

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] inorganic solid-state materials,^[20–23] high-entropy alloys,^[24,25] and bulk metallic glasses.^[26,27] 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 consists of simple supervised learning. The input variable of the model is a chemical composition, which is characterized by a descriptor vector of length 232. As detailed later, the compositional descriptor expresses the content of elements, which is conventionally operated with a predefined set of element features, such as electronegativity and atomic weight.^[28,29] The output variable is a class label corresponding to one of three structural categories: stable quasicrystal (QC), approximant crystal (AC), and “others,” which includes ordinary periodic crystals. ACs are periodic crystals composed of the similar local structural unit, such as an icosahedral cluster, as the corresponding QCs. Importantly, ACs are formed in nearby compositions of QCs, which means that their stabilization mechanisms are very similar to each other. Hence, for clarification of their common sta-

bilization mechanism, it will be of great advantage to separate ACs from all the other periodic crystals that are termed “others” here. A list of the chemical compositions of known stable 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 overall accuracy of the phase prediction task reached approximately 0.728. 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 stable 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.

2. Results

2.1. 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”

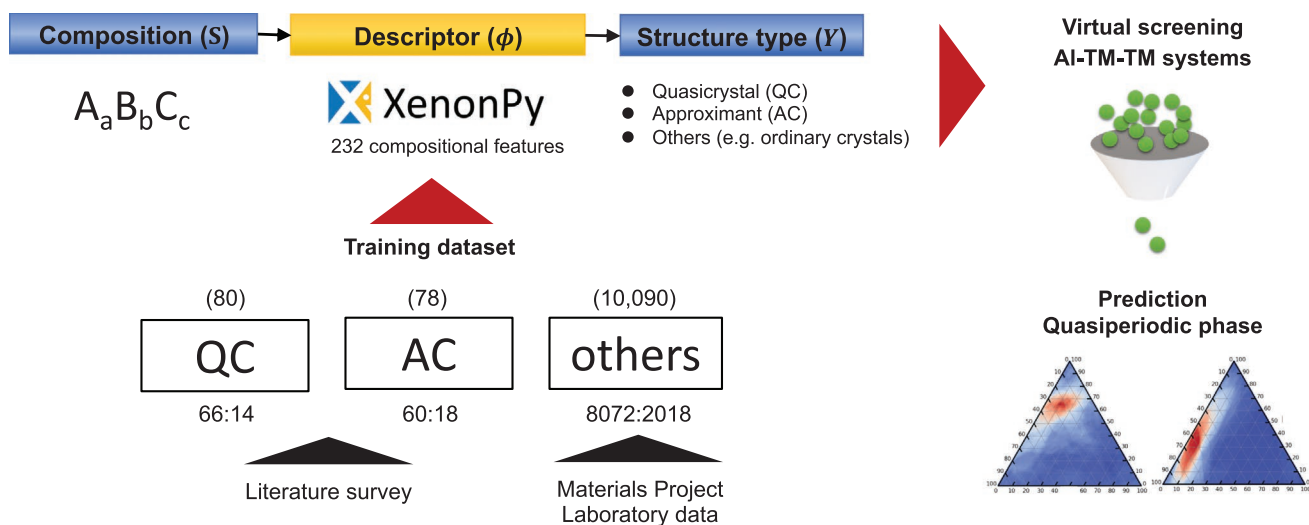


Figure 2. Machine-learning workflow. The compositional features were encoded into a 232-dimensional descriptor vector, and a prediction model was created using a random forest classifier. The trained model predicts the class label of a given chemical composition as quasicrystal (QC), approximant (AC), 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.

representing types other than the first two. We compiled a list of 80 stable quasicrystals and 78 approximants from the Crystallography of Quasicrystals handbook^[31] (see Table S1, Supporting Information and Supplementary Data (supporting file 2) for digital data). In addition, the compositions of 10 000 ordinary crystals were randomly extracted from the Materials Project database, which recorded a total of 126 335 crystals.^[32] We also used 90 crystals from our laboratory data on failed quasicrystal syntheses. These instances form the class "others". The detailed data preparation procedure is given in the Experimental 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 Experimental Section.

For each class, approximately 80% of the total data was randomly selected for training (66, 60, and 8 072 for QC, AC, and "others", respectively), and the remaining were used as a test set to measure the prediction performance (14, 18, 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 Experimental 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.

2.2. Representation of Compositional Features

Here, we describe the compositional descriptor. The chemical formula is denoted by $S = S_1^{c_1} S_2^{c_2} \dots S_K^{c_K}$. Each element of the descriptor vector of length 232 takes the form

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

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 $\eta(S^1), \dots, \eta(S^K)$ with fraction c^1, \dots, c^K were converted into the compositional feature. For f , we operated with the weighted average, weighted variance, max-pooling, and min-pooling as given by

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

Table S2, Supporting Information provides a list of the 58 element features that were implemented in XenonPy, a Python open-source platform for materials informatics that we developed.^[33] 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.

Table 1. Prediction performance for the three-class classification problem of stable 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. In Table S3, Supporting Information, we also show the performance evaluation results after eliminating the aluminum-containing compositions from the test instances.

		Predicted class				Recall	Precision	F_1
		QC	AC	Others				
True class	QC	9.63 (1.641)	3.24 (1.342)	3.13 (1.189)	QC	0.602 (0.103)	0.722 (0.090)	0.650 (0.076)
	AC	3.11 (1.555)	9.73 (1.805)	3.16 (1.573)	AC	0.608 (0.113)	0.731 (0.089)	0.658 (0.088)
	Others	0.76 (0.896)	0.42 (0.619)	2016.82 (1.024)	Others	0.999 (0.001)	0.997 (0.001)	0.998 (0.001)

2.3. Generalization Ability of the Model

We predicted the class labels of 2050 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

$$\begin{aligned} \text{Recall}(c) &= \frac{\text{TP}(c)}{\text{TP}(c) + \text{FN}(c)} \\ \text{Precision}(c) &= \frac{\text{TP}(c)}{\text{TP}(c) + \text{FP}(c)} \\ F_1(c) &= 2 \cdot \frac{\text{Recall}(c) \cdot \text{Precision}(c)}{\text{Recall}(c) + \text{Precision}(c)} \end{aligned} \quad (3)$$

$\text{TP}(c)$ denotes the number of true positives when label c is treated as positive and the other two classes as negative, and $\text{FN}(c)$ and $\text{FP}(c)$ 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. $F_1(c)$ 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 exhibit the generalized ability to identify chemical compositions that could potentially generate stable quasicrystals and approximant crystals.

2.4. Phase Prediction of Ternary Alloy Systems

Of the 100 models shown above, the model that achieved the highest prediction accuracy was selected, and 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] (TM: transition metal) 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, stable 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.

We verified the validity of the predicted phase diagrams based on the experimental stable quasicrystal and approximant phase regions of the 30 systems that were extracted from the literature.^[34–58] We found 198 papers 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 stable quasicrystal and 78 approximant compositions used for training, the remaining 30 systems were used for performance evaluation. Figure S2, Supporting Information 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.^[52] 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, $\mathcal{G}_{c_{\text{exp}}}$ denotes the set of all grid points in experimental phase $c_{\text{exp}} \in \{\text{QC}, \text{AC}, \text{others}\}$ in a diagram. Using the trained model, we calculated the mean probability $p(Y = c | \mathcal{G}_{c_{\text{exp}}})$ for each c_{exp} and $c \in \{\text{QC}, \text{AC}, \text{others}\}$ by

$$p(Y = c | \mathcal{G}_{c_{\text{exp}}}) = \frac{1}{|\mathcal{G}_{c_{\text{exp}}}|} \sum_{i \in \mathcal{G}_{c_{\text{exp}}}} p(Y_i = c | S_i) \quad (4)$$

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

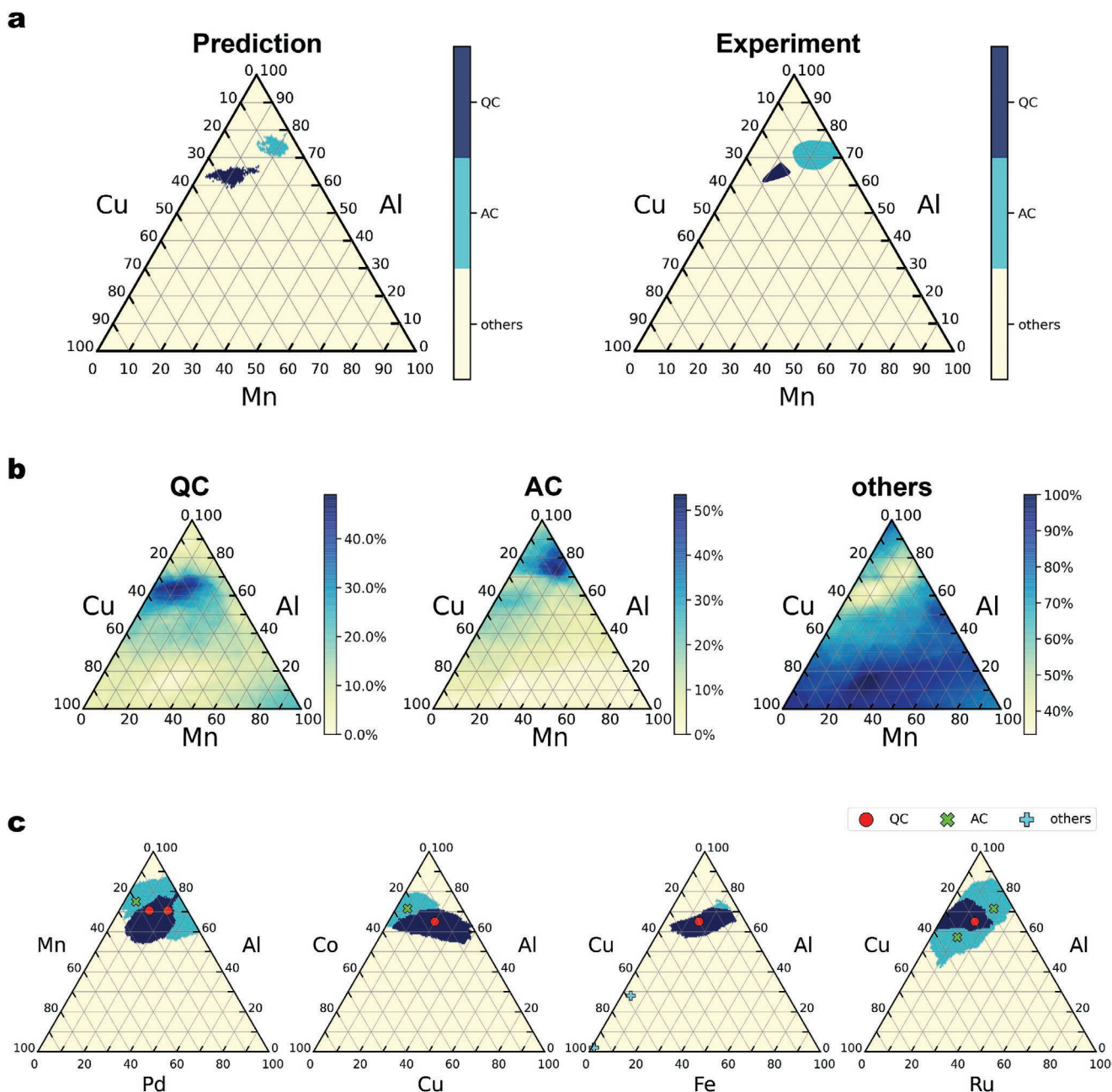


Figure 3. Phase prediction of the Al–Mn–Cu system. a) Predicted phase diagram (left) and experimental phase diagram (right) of the Al–Mn–Cu system. The three colors denote the stable quasicrystalline phase (QC), approximant phase (AC), and “others”. Despite the lack of training instances for Al–Mn–Cu, the model successfully predicts the unseen stable 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.

Table 2. Phase prediction performance for the 30 Al–TM–TM (TM: transition metal) alloy systems.

		Predicted class				Recall	Precision	F ₁
		QC	AC	Others				
True class	QC	3	1	0	QC	0.750	0.333	0.462
	AC	5	13	7	AC	0.520	0.813	0.634
	Others	1	2	27	Others	0.900	0.794	0.844

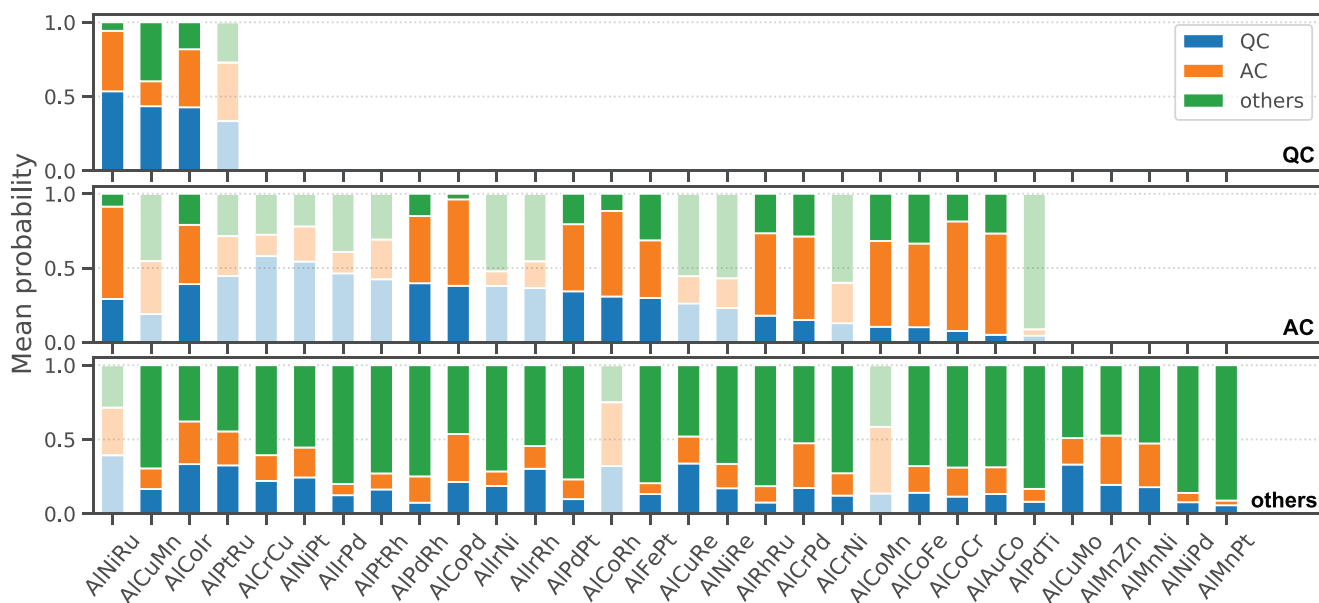


Figure 4. Prediction performance for the 30 different Al–TM–TM systems. The mean class probability was calculated in each of the experimental phase regions (top: QC, middle: AC, and bottom: “others”) using our trained random forest classifier. The bar plots shown in a transparent color represent phases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

of each class with respect to the three different phases in the 30 systems is displayed in **Figure 4**.

The results showed a similar trend to the performance metrics from the previously discussed composition-level evaluation (Table 1). The precision and recall were 0.333 and 0.750 for quasicrystals, 0.813 and 0.520 for approximants, and 0.794 and 0.900 for others, respectively. The overall prediction accuracy reached 0.728. 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, Supporting Information 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 3c). 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.

2.5. 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 stable 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 $e/a = 1.8$ is overlaid (see ref. [59] 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, Supporting Information). 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.

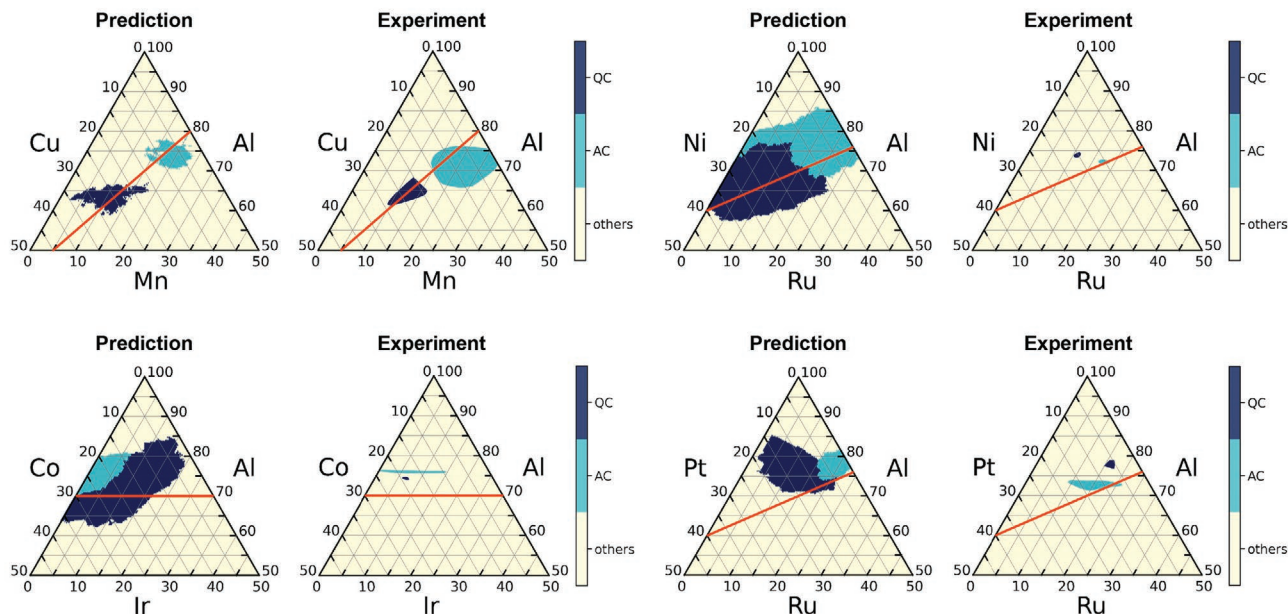


Figure 5. Predicted and experimental phase diagrams of four ternary alloy systems. The orange lines indicate the Hume-Rothery rule of valence electron concentration with $e/a = 1.8$.

2.6. 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 21925080 hypothetical compositions $\{S_i | i = 1, \dots, N\}$ ($N = 21925080$) 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.^[60] 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 h th 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, that is, the weighted average of the van der Waals radius, is consistent with the Hume-Rothery rules, where the atomic size factor is considered to con-

tribute 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 $\{QC, AC, \text{others}\}$. Descriptors with larger discrepancies in the within-class means and smaller 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 $p(Y \in \{QC, AC\} | S)$ 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 $S_{\hat{c}^1}^1 S_{\hat{c}^2}^2 S_{\hat{c}^3}^3$ 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:

$$C_h = \left\{ (\hat{c}^1, \hat{c}^2, \hat{c}^3) \left| \sum_{i=1}^3 \hat{c}^i \eta(S^i) = m_h, \sum_{i=1}^3 \hat{c}^i = 1, \hat{c}^i \geq 0 (\forall i) \right. \right\} \quad (5)$$

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, 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”),

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 ID (upper) and name (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:mendelev_number					
Mendelev'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 ⁻¹ mol ⁻¹]	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 K ⁻¹ mol ⁻¹]	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 = 0 K [eV atom ⁻¹]	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 ⁻¹ K ⁻¹]	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 K ⁻¹ mol ⁻¹]	0.34	24.44 (0.10)	24.47 (0.14)	24.81 (0.58)	24.47 (0.14)
ave:density					
Density at 295 K [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)

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* = 0K ground state calculated by density functional theory

(“ave:gs_energy”) among the highly relevant descriptors listed in Table 3. Then, we overwrote each C_i on the predicted phase diagrams for the 30 alloy systems. **Figure 6** illustrates eight selected phase diagrams (see also Figure S2, Supporting

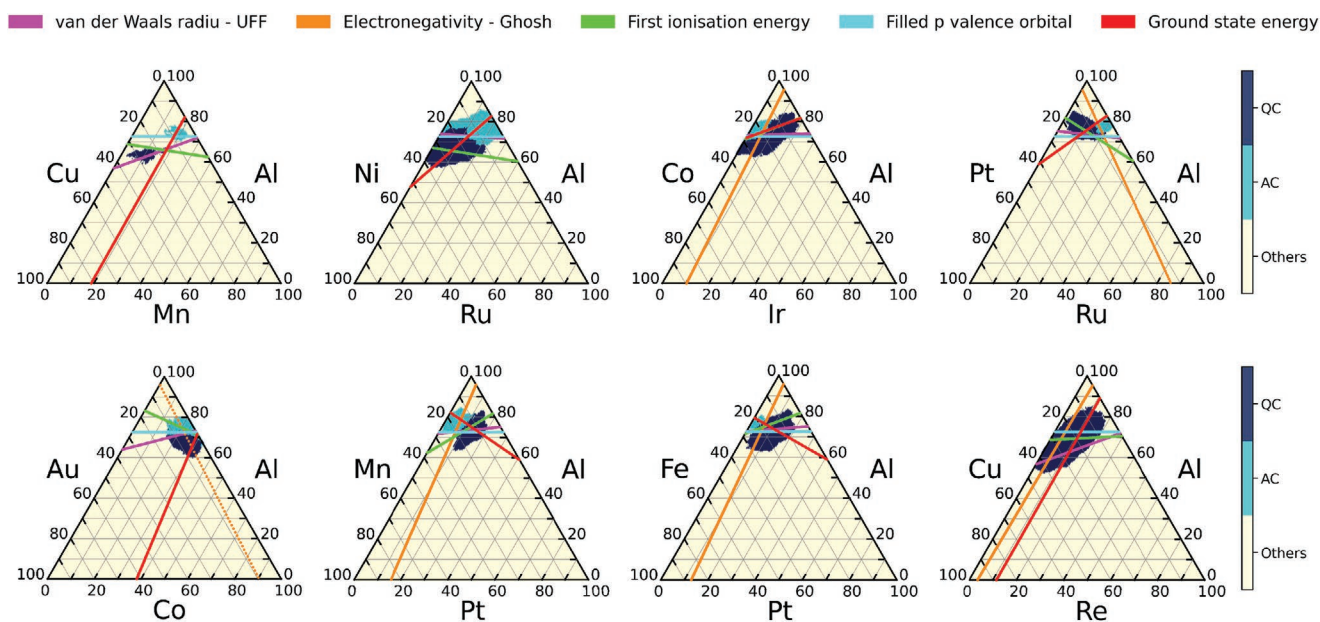


Figure 6. Five rules for the formation of QC/AC phases proposed by machine learning. The rules are represented by straight lines on the predicted phase diagrams of the eight systems. Each line represents a condition set C_h describing the weighted average of the van der Waals radius, electronegativity,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

Information for the results of all 30 systems). In almost all systems, the straight lines C_h conditioned by the five relevant descriptors passed through the predicted QC and AC phase regions. Note that each 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.

3. 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 quaternaries 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.

4. Experimental Section

Data Preparation: The list of 80 stable 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, the crystal data were downsampled by randomly extracting 10 000 instances from the overall data taken from the Materials Project database. After determining the hyperparameters based on the cross-validation as described below and training the random forest classifier, the sensitivity of the prediction accuracy was examined by varying the sample size of periodic crystals from 500 to 30 000. The result has been shown in Table S4, Supporting Information. No significant change was observed in the range from 5000 to 20 000. It should also be remarked that, instead of using the Materials Project database, other databases such as ICSD,^[61] AFLOW,^[62] and NOMAD^[63] that provide more comprehensive lists of periodic crystals can be used.

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

Compositional Pattern of Datasets: Figure 1b shows a low-dimensional representation of the compositional distribution of the data belonging to the three classes, which was used to determine the between-class difference and overlap. The compiled list of stable quasicrystals and approximants consisted of 26 binary, 120 ternary, and 12 quaternary compounds spanning 50 different elements. On the other hand, the ordinary crystal dataset consisted of unary to octonary systems with constituents spanning a broader range of elements. To more clearly visualize the difference and overlap in the class-specific distributions,

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, log₂
Bootstrap sampling in the bagging	False , True
Classification loss	Entropy , gini

only binary to quaternary crystals were shown on the plot. Furthermore, crystals containing elements other than the constituents of the stable quasicrystals and approximants were excluded. Each composition was translated 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 the 19 191 compositions were projected onto a 2D subspace using a dimensionality reduction technique called UMAP.^[30] There was no significant bias in the distribution of the three classes at the level of their constituent elements, implying that no particular combination of elements was favorable for the formation of stable quasicrystals. The visualized data pattern also suggested that previous studies on stable 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. Cross-validation was performed in the training dataset and the hyperparameters that minimized the prediction error were selected. 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 stable 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; that is, 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, models from 100 randomly selected datasets with the selected hyperparameters were also trained. Using these models, the mean and standard deviation of the performance metrics with respect to the test dataset were calculated. The learning algorithm implemented in scikit-learn^[64] v0.23.1 (<https://github.com/scikit-learn/scikit-learn/releases/tag/0.23.1>) was employed to train the models.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was supported in part by a MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number 19H05820). R.Y. acknowledges the financial support from a JSPS Grant-in-Aid for Scientific Research (A) 19H01132 from the Japan Society for the Promotion of Science (JSPS), JST CREST Grant Number JPMJCR19I3, and JPNP16010 commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

R.Y. and K.K. designed the conceptual idea and proof outline. C.L. and R.Y. wrote the manuscript and carried out the data analysis. E.F., Y.K., Y.I., A.I., R.T., and K.K. worked out the collection and curation of the dataset. All authors discussed the results and commented on the manuscript.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

approximant crystals, high-throughput screening, machine learning, materials informatics, quasicrystals

Received: April 1, 2021
Revised: April 30, 2021
Published online: July 19, 2021

- [1] D. Shechtman, I. Blech, D. Gratias, J. W. Cahn, *Phys. Rev. Lett.* **1984**, 53, 1951.
- [2] A. P. Tsai, A. Inoue, T. Masumoto, *Jpn. J. Appl. Phys.* **1987**, 26, L1505.
- [3] A. P. Tsai, A. Inoue, T. Masumoto, *Mater. Trans. JIM* **1989**, 30, 463.
- [4] A. P. Tsai, A. Inoue, Y. Yokoyama, T. Masumoto, *Mater. Trans. JIM* **1990**, 31, 98.
- [5] H. Takakura, C. P. Gómez, A. Yamamoto, M. De Boissieu, A. P. Tsai, *Nat. Mater.* **2007**, 6, 58.
- [6] T. Yamada, H. Takakura, M. de Boissieu, A. P. Tsai, *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2017**, 73, 1125.
- [7] K. Kimura, H. Iwahashi, T. Hashimoto, S. Takeuchi, U. Mizutani, S. Ohashi, G. Itoh, *J. Phys. Soc. Japan* **1989**, 58, 2472.
- [8] U. Mizutani, Y. Sakabe, T. Shibuya, K. Kishi, K. Kimura, S. Takeuchi, *J. Phys. Condens. Matter* **1990**, 2, 6169.
- [9] H. Akiyama, Y. Honda, T. Hashimoto, K. Edagawa, S. Takeuchi, *Jpn. J. Appl. Phys.* **1993**, 32, Part 2, No. 7B L1003.
- [10] T. Watanuki, S. Kashimoto, D. Kawana, T. Yamazaki, A. Machida, Y. Tanaka, T. J. Sato, *Phys. Rev. B* **2012**, 86, 094201.
- [11] K. Deguchi, S. Matsukawa, N. K. Sato, T. Hattori, K. Ishida, H. Takakura, T. Ishimasa, *Nat. Mater.* **2012**, 11, 1013.
- [12] K. Kamiya, T. Takeuchi, N. Kabeya, N. Wada, T. Ishimasa, A. Ochiai, K. Deguchi, K. Imura, N. K. Sato, *Nat. Commun.* **2018**, 9, 154.
- [13] W. Hume-Rothery, *J. Inst. Metals* **1926**, 35, 295.
- [14] A. P. Tsai, *Sci. Technol. Adv. Mater.* **2008**, 9, 013008.
- [15] K. Nakayama, A. Mizutani, Y. Koyama, *J. Phys. Soc. Japan* **2016**, 85, 114602.
- [16] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, *Nat. Mater.* **2016**, 15, 1120.
- [17] E. C. Hansen, D. J. Pedro, A. C. Wotal, N. J. Gower, J. D. Nelson, S. Caron, D. J. Weix, *Nat. Chem.* **2016**, 8, 1126.
- [18] M. Sumita, X. Yang, S. Ishihara, R. Tamura, K. Tsuda, *ACS Cent. Sci.* **2018**, 4, 1126.
- [19] S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, R. Yoshida, *npj Comput. Mater.* **2019**, 5, 66.
- [20] A. O. Olynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, A. Mar, *Chem. Mater.* **2016**, 28, 7324.
- [21] R. Matsumoto, Z. Hou, H. Hara, S. Adachi, H. Takeya, T. Irifune, K. Terakura, Y. Takano, *Appl. Phys. Express* **2018**, 11, 093101.
- [22] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, *Phys. Rev. Lett.* **2015**, 115, 205901.
- [23] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, *Phys. Rev. X* **2014**, 4, 011019.
- [24] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, *Acta Mater.* **2019**, 170, 109.
- [25] Y. Zhang, C. Wen, C. Wang, S. Antonov, D. Xue, Y. Bai, Y. Su, *Acta Mater.* **2020**, 185, 528.
- [26] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, *Sci. Adv.* **2018**, 4, eaaq1566.
- [27] J. Xiong, S.-Q. Shi, T.-Y. Zhang, *Mater. Des.* **2020**, 187, 108378.
- [28] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, 2, 16028.
- [29] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **2017**, 95, 144110.
- [30] L. McInnes, J. Healy, N. Saul, L. Großberger, *J. Open Source Softw.* **2018**, 3, 861.
- [31] W. Steurer, S. Deloudi, *Crystallography of Quasicrystals: Concepts, Methods and Structures*, Springer Series in Materials Science, vol. 126, Springer, Berlin/Heidelberg, Germany **2009**.
- [32] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, 1, 011002.
- [33] XenonPy Platform, <https://github.com/yoshida-lab/XenonPy> (accessed: April 2021).
- [34] S. Mi, B. Grushko, C. Dong, K. Urban, *J. Alloys Compd.* **2003**, 354, 148.
- [35] B. Grushko, T. Velikanova, *Comput. Coupling Phase Diagrams Thermochem.* **2007**, 31, 217.
- [36] D. Pavlyuchkov, B. Grushko, T. Velikanova, *J. Alloys Compd.* **2008**, 453, 191.
- [37] B. Grushko, W. Kowalski, D. Pavlyuchkov, S. Mi, M. Surowiec, *J. Alloys Compd.* **2009**, 485, 132.
- [38] B. Grushko, W. Kowalski, S. Balanetsky, *J. Alloys Compd.* **2009**, 479, L59.
- [39] B. Grushko, W. Kowalski, M. Surowiec, *J. Alloys Compd.* **2010**, 491, L5.
- [40] W. Kowalski, B. Grushko, D. Pavlyuchkov, M. Surowiec, *J. Alloys Compd.* **2010**, 496, 129.
- [41] D. Kapush, T. Velikanova, B. Grushko, *J. Alloys Compd.* **2010**, 497, 105.
- [42] B. Grushko, D. Kapush, T. Y. Velikanova, S. Samuha, L. Meshi, *J. Alloys Compd.* **2011**, 509, 8018.
- [43] B. Grushko, S. Mi, *J. Alloys Compd.* **2011**, 509, L30.
- [44] B. Grushko, D. Kapush, V. Konoval, V. Shemet, *Powder Metall. Met. Ceram.* **2011**, 50, 462.
- [45] S. Balanetsky, G. Meisterernst, B. Grushko, M. Feuerbacher, *J. Alloys Compd.* **2011**, 509, 3795.
- [46] O. V. Zaikina, V. G. Khorujaya, D. Pavlyuchkov, B. Grushko, T. Y. Velikanova, *J. Alloys Compd.* **2011**, 509, 43.
- [47] B. Grushko, D. Kapush, L. Meshi, *J. Alloys Compd.* **2012**, 514, 60.
- [48] B. Grushko, D. Kapush, S. Samuha, L. Meshi, *J. Alloys Compd.* **2014**, 600, 125.
- [49] B. Grushko, *J. Alloys Compd.* **2015**, 636, 329.
- [50] D. Kapush, S. Samuha, L. Meshi, T. Y. Velikanova, B. Grushko, *J. Phase Equilibria Diffus.* **2015**, 36, 327.
- [51] B. Grushko, D. Pavlyuchkov, S. Mi, S. Balanetsky, *J. Alloys Compd.* **2016**, 677, 148.

- [52] B. Grushko, S. B. Mi, *J. Alloys Compd.* **2016**, 688, 957.
- [53] S. Samuha, B. Grushko, L. Meshi, *J. Alloys Compd.* **2016**, 670, 18.
- [54] B. Grushko, *J. Alloys Compd.* **2017**, 729, 426.
- [55] B. Grushko, W. Kowalski, S. Mi, *J. Alloys Compd.* **2018**, 739, 280.
- [56] B. Grushko, *J. Alloys Compd.* **2019**, 772, 399.
- [57] B. Grushko, *J. Alloys Compd.* **2019**, 792, 1223.
- [58] B. Grushko, *J. Alloys Compd.* **2020**, 829, 154444.
- [59] K. Kitahara, K. Kimura, *Z. Kristallogr. - Cryst. Mater.* **2017**, 232, 507.
- [60] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, P. C. Sabeti, *Science* **2011**, 334, 1518.
- [61] A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, *Acta Cryst. B* **2002**, 58, 364.
- [62] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, 58, 218.
- [63] C. Draxl, M. Scheffler, *J. Phys. Mater.* **2019**, 2, 036001.
- [64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, 12, 2825.