






## Representation of materials by kernel mean embedding

Minoru Kusaba <sup>1,\*</sup>,† Yoshihiro Hayashi <sup>1,2,\*</sup> Chang Liu <sup>1</sup> Araki Wakiuchi <sup>3,4</sup> and Ryo Yoshida <sup>1,2,5,‡</sup><sup>1</sup>Research Organization of Information and Systems, The Institute of Statistical Mathematics, Tachikawa 190-8562, Japan<sup>2</sup>SOKENDAI, The Graduate Institute for Advanced Studies, Tachikawa 190-8562, Japan<sup>3</sup>JSR Corporation, 3-103-9 Tonomachi, Kawasaki-ku, Kawasaki, Kanagawa 210-0821, Japan<sup>4</sup>Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan<sup>5</sup>National Institute for Materials Science, Research Network and Facility Services Division, Tsukuba 305-0047, Japan

(Received 2 June 2023; revised 1 September 2023; accepted 6 September 2023; published 16 October 2023)

For using machine learning to predict material properties, the feature representation of the materials given to the model plays a fundamental role. A model describes material properties as a function of any given material system expressed as a fixed-length numeric vector, often called a descriptor. However, in most cases, the variables of interest are nontrivial for encoding their compositional or structural features, such as molecules, crystal systems, chemical compositions, and composite materials, into a fixed-length vector. Conventionally, to translate such a multicomponent system into a fixed-length vector, the distribution of predefined component features is summarized into a few summary statistics. The disadvantage of this reduction operation is that some distributional information, such as multimodality, is lost in the vectorization process. Here, we present a general class of material descriptors motivated by the machine-learning theory of kernel mean embedding. Unlike conventional descriptors, kernel mean embedding can retain all information regarding the distribution of component features in the vectorization process. Furthermore, the kernel mean descriptor uniquely determines the inverse map to the original material space. We demonstrate the expressive power and versatility of the kernel mean descriptor in various applications, including the prediction of the formation energy of inorganic compounds, prediction of the chemical composition to form quasicrystalline materials, and the use of force-field parameters to characterize polymeric materials.

DOI: [10.1103/PhysRevB.108.134107](https://doi.org/10.1103/PhysRevB.108.134107)

## I. INTRODUCTION

Predicting material properties using machine learning has the potential to significantly accelerate the discovery of innovative materials. Machine-learning models enable rapid high-throughput virtual screening across millions or billions of candidate materials spanning an enormous search space [1–6]. Generally, a model describes physicochemical, electronic, thermodynamic, or mechanical properties as a function of the input materials, which are given in various forms, such as small or macromolecules, crystal systems, chemical or raw material compositions, or their mixtures. To put the problem into a machine-learning framework, such a nonnumeric variable must be transformed into a fixed-length numeric vector called a descriptor, which represents the compositional or structural features of the given material [7–19]. A model is trained on a given data set to learn a mapping from the vectorized features to the target properties. In this workflow,

the feature representation of the input materials is the key to boosting predictive power.

The objects represented include molecules, compositional inputs, crystal structures, and composite materials. A class of descriptors called molecular fingerprints represents a chemical structure by vectorizing the presence or absence, or the number of each substructure, given hundreds or thousands of chemical fragments [11–16]. Another descriptor class includes quantitative descriptors, representing the topological and geometrical features or physicochemical properties of molecular systems [20–26]. Compositional descriptors express the number of chemical elements or raw materials. For example, a conventional compositional descriptor operates with a predefined set of element features such as electronegativity and atomic weight [7,27,28]. With a given composition ratio, the feature values of the constituent elements are collapsed into a quantity that describes a compositional feature; for example, using the weighted mean and weighted variance of the element features. A crystal structure is generally characterized by encoding the local structural environment of each atom in a unit cell into a set of quantities to define a structural descriptor [8–10,29,30]. Similar to a compositional descriptor, the local atomistic features in a crystal system are reduced to summary statistics. In recent years, there has been an increasing trend of representing the structure of materials as graphs and predicting their properties using graph neural networks [31–33]. A natural representation of chemical structures is labeled as an undirected graph. The periodic configuration of

\*These authors contributed equally to this work.

†Corresponding author: kusaba@ism.ac.jp

‡Corresponding author: yoshidar@ism.ac.jp

atoms in a crystal system can also be translated into a crystal graph, which represents the neighboring relations of atoms in infinitely arranged unit cells [31]. Generally, the component features of the atoms and atom groups that define a graph are vectorized via graph convolution operations learned from a given training data set.

As described above, common descriptors collapse the component features of building blocks in a given material system into a set of quantities by taking the mean, variance, etc. Materials are inherently nontrivial to encode into fixed-length vectors. This is because a material system comprises varying numbers of building blocks. For example, the number of elements differs between binary and ternary compounds. In this case, a descriptor is designed to reduce element features (e.g., atomic weights) to a summary statistic by calculating their weighted means for a given composition ratio [7]. For the vector representation of polymers based on force-field parameters, attribute values are assigned to atoms or atom groups consisting of bonds and dihedral angles [34]. It is then necessary to reduce the component features, whose numbers vary across different polymers, to a descriptor vector of the same dimension. Furthermore, in some cases, a designed descriptor is required to hold invariance to the exchange of components or building blocks. For example, there is no ordering relationship between the chemical elements in a compound; therefore, the descriptors must be invariant to their exchange.

This paper presents a general class of material descriptors that relies on the machine-learning theory of kernel mean embedding [35]. The proposed method treats the per-building block features in a material as samples from a probability distribution. One of the drawbacks of the existing methods is that some essential features of the probability distribution, such as higher-order moments and multimodality, are lost when collapsing into a finite set of summary statistics. Instead, kernel mean embedding with a specific class of kernel functions, called characteristic kernels, can uniquely preserve all the information about the probability distribution, which is mapped to a feature space called the reproducing kernel Hilbert space (RKHS) [36]. In machine learning, kernel mean embedding provides a theoretical basis for supervised learning, in which probability distributions are treated as model inputs [37]. Following this framework, we establish a general methodology for kernel mean descriptors encompassing various material objects. Furthermore, we propose a method for inverse translation from descriptors to materials that takes advantage of the linearity of kernel mean embedding with respect to the component ratio. We also prove that the kernel mean descriptor uniquely determines the inverse map to the original material space. We demonstrate the expressive power and versatility of kernel mean descriptors in various applications, such as the energy prediction of inorganic compounds, prediction of the chemical composition to form quasicrystalline materials [27], and representation of compositional and structural features of polymers using force-field parameters in an empirical potential energy function.

## II. METHODS

### A. Preliminaries

A material system  $X$  is expressed as a collection of  $N_X$  constitutional elements or building blocks,  $x_1, \dots, x_{N_X}$ , with

their contents  $w_1, \dots, w_{N_X}$  normalized such that they are non-negative and sum to one. Each component is characterized by a feature  $\lambda_i = \lambda(x_i) \in \mathbb{R}$  ( $i = 1, \dots, N_X$ ), which is hereafter referred to as a component feature. In general, an element in a descriptor vector  $\phi_{(f,\lambda)}(X) \in \mathbb{R}$  for the entire system  $X$  can be expressed as

$$\phi_{(f,\lambda)}(X) = f(w_1, \dots, w_{N_X}, \lambda(x_1), \dots, \lambda(x_{N_X})) \quad \forall \lambda \in \Lambda, f \in \mathcal{F}. \quad (1)$$

A combination of component descriptors  $\lambda \in \Lambda$  and summary functions  $f \in \mathcal{F}$  constitute the descriptor vector.

In many applications, as exemplified later, the number of components  $N_X$  varies across different  $X$ . In addition,  $x_1, \dots, x_{N_X}$  should be treated as a set variable or sample from a probability distribution, as well as  $\{(w_i, \lambda_i) | i = 1, \dots, N_X\}$ . This implies that the exchange in any pair of  $x_1, \dots, x_{N_X}$  should not alter the encoding. To handle such variable-length objects and maintain exchange invariance, a conventional descriptor employs summary statistics for  $f$  to collapse  $\{(w_i, \lambda_i) | i = 1, \dots, N_X\}$  into a scalar quantity  $\phi_{(f,\lambda)}(X)$ . Commonly used summary statistics include the weighted mean, weighted variance, max-pooling, and min-pooling, as follows:

$$\begin{aligned} \phi_{(\text{mean},\lambda)}(X) &= \sum_{i=1}^{N_X} w_i \lambda_i, \\ \phi_{(\text{var},\lambda)}(X) &= \sum_{i=1}^{N_X} w_i (\lambda_i - \phi_{(\text{mean},\lambda)}(X))^2, \\ \phi_{(\text{max},\lambda)}(X) &= \max\{\lambda_1, \dots, \lambda_{N_X}\}, \\ \phi_{(\text{min},\lambda)}(X) &= \min\{\lambda_1, \dots, \lambda_{N_X}\}. \end{aligned} \quad (2)$$

Higher-order moments, such as skewness and kurtosis, can also be used [38].

Generally, descriptor design aims to characterize the probability distribution. In other words, a histogram consisting of  $N_X$  sample points and their probability masses  $\{(w_i, \lambda_i) | i = 1, \dots, N_X\}$  can be interpreted as an approximation of the probability distribution using an empirical distribution or the probability mass function itself. From this perspective, there is a risk of losing important features of the original probability distribution through the reduction operation to a few summary statistics. For example, the distribution of component features can be highly multimodal, and the multimodal nature of the distribution may be a dominant factor in determining physicochemical properties. However, moment statistics of up to the third or fourth order cannot preserve the features related to the multimodality of the distribution, as illustrated in the right panel in Fig. 1.

### B. Kernel mean descriptors

Let  $P_X(\lambda) \in \mathcal{P}$  be a probability distribution function followed by  $N_X$  component features  $\lambda_i = \lambda(x_i)$  ( $i = 1, \dots, N_X$ ) of the constituent elements  $x_1, \dots, x_{N_X}$  in material  $X$ . Here,  $\mathcal{P}$  denotes the set of probability measures. The objective is to represent the distribution  $P_X$  of  $X$  as a fixed-length vector. Kernel means embedding is a technique for mapping any

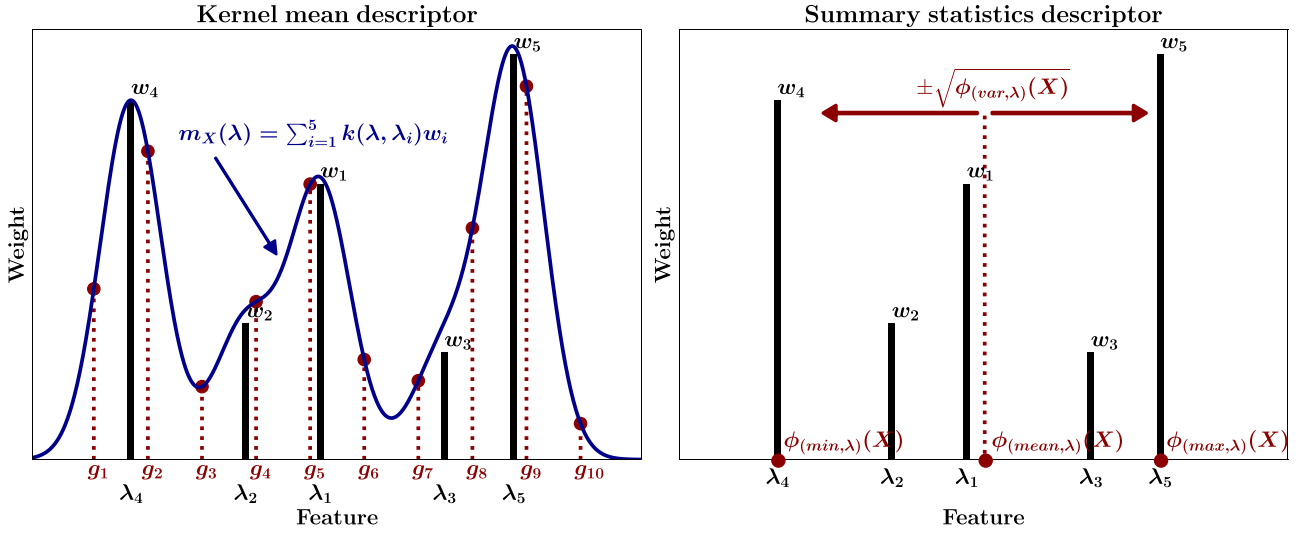


FIG. 1. Schematic view of the reduction operation using kernel mean embedding (left) and summary statistics (right). In both panels, the black bars indicate the probability mass function  $\{(w_i, \lambda_i) | i = 1, \dots, 5\}$ , which represents a material system  $X$ . Left: The blue curve shows the kernel mean  $m_X(\lambda)$ . The kernel mean descriptor  $(m_X(g_1), \dots, m_X(g_{10}))^\top$  is obtained by discretizing  $m_X(\lambda)$  at the ten grid points  $(g_1, \dots, g_{10})$  on the feature space. Right: The three red dots show the locations of the weighted mean  $\phi_{(\text{mean}, \lambda)}(X)$ , max-pooling  $\phi_{(\text{max}, \lambda)}(X)$ , and min-pooling  $\phi_{(\text{min}, \lambda)}(X)$ , respectively. The red arrow indicates the weighted variance  $\phi_{(\text{var}, \lambda)}(X)$  [the arrow length is set to  $\sqrt{\phi_{(\text{var}, \lambda)}(X)}$ ].

given probability distribution  $P_X(\lambda)$  onto its embedding  $m_X(\cdot)$  in an RKHS  $\mathcal{H}_k$ , defined by a kernel function  $k$ . The kernel  $k(\lambda, \lambda')$  is a bivariate symmetric function that holds specific properties with positive definite and reproducibility. When one argument is fixed at  $\lambda'$ , the univariate function  $k(\cdot, \lambda')$  becomes an element of  $\mathcal{H}_k$ . We do not present further details of the positive definite kernel here (e.g., see Ref. [39]).

The kernel mean embedding  $m_X(\cdot)$  of  $P_X$  is defined as follows:

$$m_X(\cdot) := \mathbb{E}_{\lambda \sim P_X} [k(\cdot, \lambda)] = \int_{S_\lambda} k(\cdot, \lambda) dP_X(\lambda). \quad (3)$$

The integral over the support  $S_\lambda$  of the component feature expresses the expected value of kernel  $k$  with respect to the probability measure  $P_X(\lambda)$ . When the kernel is positive definite, kernel mean  $m_X(\cdot)$  belongs to  $\mathcal{H}_k$ .

Here, we introduce a class of positive-definite kernels called characteristic kernels [40]. The map  $\mathcal{P} \rightarrow \mathcal{H}_k$  ( $P_X \mapsto m_X(\cdot)$ ), defined by the mean operation with  $k$  belonging to this class, is known to be injective. This implies the following:

$$m_X(\cdot) = m_Y(\cdot) \Leftrightarrow P_X(\cdot) = P_Y(\cdot). \quad (4)$$

In other words, any probability distributions  $P_X$  and  $P_Y$  are the same if their kernel mean embeddings are the same. A given  $m_X(\cdot)$  of a characteristic kernel can uniquely determine the respective probability distribution, implying that the complete information of any  $P_X \in \mathcal{P}$  (i.e., all-moment statistics) can be retained in the embedding space. The Gaussian and Laplace radial basis function (RBF) kernels are well-known characteristic kernels in Euclidean space [41]. Owing to this property and ease of handling, the Gaussian RBF kernel  $k(\lambda, \lambda') = \frac{1}{Z} \exp(-\frac{(\lambda - \lambda')^2}{2\sigma^2})$  was employed as the default kernel function in this study, where  $Z$  denotes the normalizing constant.

Generally, a material system is defined by a finite number of components. Hence, the domain of the probability

distribution  $P_X(\lambda)$  is defined in the discrete set. Therefore, the probability distribution function can be expressed by the probability mass function as

$$P_X(\lambda) = \sum_{i=1}^{N_X} \delta(\lambda - \lambda_i) w_i, \quad (5)$$

where  $\delta(\lambda - \lambda_i)$  is the Dirac delta function, which takes the value of one if  $\lambda = \lambda_i$  and zero otherwise. In this case, the kernel mean is given by

$$m_X(\lambda) = \sum_{i=1}^{N_X} k(\lambda, \lambda_i) w_i. \quad (6)$$

For example, using the Gaussian RBF kernel function normalized such that the integral over the overall domain is equal to one [ $\int_{S_\lambda} k(\lambda, \lambda') d\lambda = 1$  ( $\forall \lambda'$ )], the kernel mean becomes the density function of the Gaussian mixture

$$m_X(\lambda) = \frac{1}{\sqrt{2\pi}\sigma^2} \sum_{i=1}^{N_X} \exp\left(-\frac{(\lambda - \lambda_i)^2}{2\sigma^2}\right) w_i, \quad (7)$$

where  $\sigma^2$  denotes the variance parameter, which is assumed to be homogeneous concerning  $N_X$  component distributions. Notably, although the component features are essentially discrete variables, the domain of the kernel mean function is a continuous set.

The kernel mean embedding is equivalent to kernel density estimation [42] for the weighted sample set  $\mathcal{D} = \{(w_i, \lambda_i) | i = 1, \dots, N_X\}$ , except for a normalization constant. In other words, it smoothens the histogram  $\mathcal{D}$ . The variance  $\sigma^2$  of the Gaussian RBF kernel function is a hyperparameter that controls smoothness. The larger the variance, the smoother the estimated density function. Conversely, the variance approaches zero and it converges to the original discrete distribution in Eq. (5), that is, the histogram  $\mathcal{D}$ .

The kernel mean descriptor directly vectorizes the shape of the probability distribution function followed by the component features. In contrast, the descriptors based on summary statistics in Eq. (2) reduce the distributional features to a finite set of moment statistics, such as the mean and variance. This reduction results in the loss of important elements in the probability distribution. For example, for the summary statistics descriptor, there are many compositions in which the weighted means of the element features have the same value. In such cases, redundancies occur when solving an inverse problem. In other words, the resulting machine learning models cannot distinguish between different compositions that have the same weighted mean. The kernel mean descriptor can solve such an ill-conditioned problem by retaining almost complete information regarding the probability distribution.

A schematic of the reduction operation using kernel mean embedding and summary statistics is presented in Fig. 1. The right panel indicates that the multimodality of the distribution cannot be captured by the four summary statistics in Eq. (2). However, as shown in the left panel, the kernel mean descriptor successfully captures the three peaks of the distribution and the valleys between them.

As previously mentioned, the kernel mean  $m_X(\lambda)$  is a continuous function of  $\lambda$  (a vector of infinite dimensions). Therefore, except for special cases, such as kernel regression [43] or support vector machines [44],  $m_X(\lambda)$  must be discretized for use as an input variable in ordinary supervised learning. In this paper, a descriptor vector for  $\lambda$  is given by  $\phi_\lambda(X) = (m_X(g_1), \dots, m_X(g_d))$ , which is discretized at  $d$  equally spaced grid points  $(g_1, \dots, g_d)$  between the predefined maximum and minimum values. The dimension  $d$  of the descriptor and the smoothing parameter  $\sigma^2$  of the Gaussian RBF kernel form the hyperparameters, which can be automatically adjusted according to the problem setting or be optimized during the cross-validation process in supervised learning. Finally, the descriptors for the various component features  $\lambda \in \Lambda$  are concatenated to define the overall descriptor  $\phi(X) = (\phi_\lambda(X))_{\lambda \in \Lambda}$ . Therefore, if the dimension of the component descriptor is  $K$  [i.e.,  $n(\Lambda) = K$ ], then the material system  $X$  is represented as a vector of  $d \times K$  dimensions using the kernel mean descriptor.

Unlike the summary statistics descriptor, the dimension of the kernel mean descriptor can be tailored to the problem setting by arbitrarily changing the number of grid points  $d$ , which is a great advantage in practice. In particular, this feature helps prevent overfitting problems when the amount of data is limited. The effect of changing the dimension of the kernel mean descriptor on prediction performances is investigated in the Results section.

### C. Inverse translation of the kernel mean descriptors

The inverse translation of a descriptor  $\phi(X)$  is a task of identifying a material system  $X$  for which  $\phi(X)$  takes a specific value. If material  $X$  is a mixture system, the task is equivalent to estimating the constitutional elements of  $X$  and their component ratios (weights),  $\{(w_i, x_i) | i = 1, \dots, N_X\}$ , from a given  $\phi(X)$ . For the kernel mean embedding, the linearity with respect to the weights [Eq. (6)] can be used to

establish a general framework for the inverse translation of the kernel mean descriptors with quadratic programming.

Suppose  $\{x_1, \dots, x_N\}$  is a set of all possible constitutional elements for a material system  $X$  and call it a component set. Let  $\{w_1, \dots, w_N\}$  be the weights on the component set. Each constitutional element  $x_i$  is characterized by  $K$  component features as  $\lambda_i^k = \lambda^k(x_i) \in \mathbb{R}$  ( $k = 1, \dots, K$ ), and  $d$  grid points  $(g_1^k, \dots, g_d^k)$  are defined for each feature  $\lambda^k$ . With the  $N \times d$  matrix  $G^k$  whose elements are given as  $G_{ij}^k = k(\lambda_i^k, g_j^k)$ , the kernel mean descriptor for the  $k$ th feature  $\phi_{\lambda^k}(X)$  is expressed as

$$\phi_{\lambda^k}(X) = G^{k\top} w, \quad (8)$$

where  $w = (w_1, \dots, w_N)^\top$ . Therefore, the overall descriptor  $\phi(X)$  can be summarized as follows:

$$\phi(X) = \begin{pmatrix} \phi_{\lambda^1}(X) \\ \vdots \\ \phi_{\lambda^K}(X) \end{pmatrix} = \begin{pmatrix} G^{1\top} \\ \vdots \\ G^{K\top} \end{pmatrix} w = Hw. \quad (9)$$

The  $dK \times N$  matrix  $H$  is given. The task of inverse translation is to estimate the unknown weights  $w = w^*$  that minimize the discrepancy  $\|\phi^* - Hw\|^2$  for any given  $\phi^*$ . The objective function of this minimization problem is expressed as follows:

$$\begin{aligned} & \min_w \|\phi^* - Hw\|^2 \\ & \text{such that } \mathbf{1}^\top w = 1, \\ & w \geq \mathbf{0}, \end{aligned} \quad (10)$$

where  $\mathbf{1}$  and  $\mathbf{0}$  denote  $N$ -dimensional vectors of 1 and 0, respectively. The linear constraints on  $w$  ensure that  $w$  is non-negative, or, more specifically, that it is a probability vector. Here, the optimization problem of Eq. (10) is rewritten as follows:

$$\begin{aligned} & \min_w \frac{1}{2} w^\top H^\top H w - \phi^{*\top} H w \\ & \text{such that } \mathbf{1}^\top w = 1, \\ & w \geq \mathbf{0}. \end{aligned} \quad (11)$$

Here, if the matrix  $H^\top H$  is full rank (i.e.,  $\text{rank}(H^\top H) = N$ ), the objective function is strictly convex, so there is a unique optimal solution. This means that if  $H^\top H$  is full rank, any kernel mean descriptor  $\phi^*$  is guaranteed to be uniquely mapped to a particular material system  $X^*$  with the unique solution  $w^*$ . For the matrix  $H^\top H$  to be full rank, the rank of the  $dK \times N$  matrix  $H$  must be  $N$ . Here, note that  $d$ , the number of grid points, is a user-adjustable value; thus, the matrix  $H^\top H$  can be controlled to be full rank by increasing  $d$  until  $\text{rank}(H) = N$  is satisfied. As a solver for Eq. (11), we used the quadprog library [45] in Python, which is designed for solving a strictly convex quadratic program with the Goldfarb and Idnani dual algorithm [46]. This inverse translation algorithm is implemented in the Python code of the kernel mean descriptor, which is available on GitHub [47].

## III. APPLICATIONS

This study applies the kernel mean descriptor to three examples, which are detailed in the following subsections.



### A. Formation energy prediction of inorganic compounds

Chemical composition is defined as a set of elements  $\{x_1, \dots, x_{N_X}\}$  and their contents  $\{w_1, \dots, w_{N_X}\}$ . Various element features or per-element physicochemical quantities have been considered component features. Table I lists the 58 element features implemented in XENONPY, an open-source Python platform for material informatics that we developed [27,48,49]. The element feature set includes the atomic number, bond radius, van der Waals (vdW) radius, electronegativity, thermal conductivity, band gap, polarizability, boiling point, melting point, and number of valence electrons in each orbital.

The data set used for this experiment was obtained from the Materials Project [50]. Among all the inorganic compounds (146 323) in the Materials Project, we selected all the stable compounds (energy above hull = 0). The data set consists of the formation energies per atom of 35 463 inorganic compounds that were obtained from first-principles calculations. The chemical compositions of the 35 463 compounds consist of elements with atomic numbers 1–94 (i.e., H–Pu), where the 58 element features listed in Table I are fully available. The number of elements for each compound varied from one to seven.

Using the above 58 element features ( $K = 58$ ), we vectorized the chemical composition using the summary statistics descriptor and kernel mean descriptor. Subsequently, with each descriptor, we constructed a machine-learning model that predicts the formation energy (eV/atom) of stable inorganic compounds.

For the summary statistics descriptors, we used four summary statistics: weighted mean, weighted variance, max-pooling, and min-pooling, as described in Eq. (2). Therefore, the summary statistics descriptor represents composition  $X$  as a vector of length 232 ( $= 4 \times 58$ ).

To generate the kernel mean descriptor, for each element feature space  $\lambda^k$  ( $k = 1, \dots, 58$ ), the kernel mean  $m_X(\lambda)$  was discretized at the  $d$  equally spaced grid points  $\{g_1^k, \dots, g_d^k\}$  between a maximum and minimum values of the component set (i.e.,  $g_1^k$  and  $g_d^k$  are set to be the minimum and maximum values of  $\{\lambda^k(x_1), \dots, \lambda^k(x_{94})\}$ , respectively). Here, the number of grids  $d$  was set to 10. Therefore, the kernel mean descriptor represents the composition  $X$  using a vector of length 580 ( $= 10 \times 58$ ). In this experiment, the smoothing parameter  $\sigma^2$  was set to  $\sigma^2 = \frac{|\log_2 - g_1^k|^2}{2}$  for each feature space; thus, the sum of the kernel mean descriptor  $\sum_{i=1}^d m_X(g_i^k)$  was constant (approximately  $\sqrt{\pi}$ ) for any given  $d$  and  $\lambda^k$ . In Fig. 2(b), the kernel mean descriptors and summary statistics descriptors of the 35 463 compounds, color coded by their formation energies, are visualized onto the two-dimensional manifold of principal component analysis (PCA).

### B. Prediction of chemical composition to form quasicrystalline phase

Quasicrystals (QCs) have emerged as a third class of solid-state materials, distinguished from periodic crystals and amorphous solids that do not have the translational symmetry of ordinary crystals but have a high degree of order in their atomic arrangement. Since the first QC was discovered in 1984 [52], approximately 100 stable QCs have been discov-

ered to date. However, in recent years, the pace of discovery of QCs has slowed significantly, possibly because of the lack of clear design guidelines for synthesizing stable QCs. To accelerate the discovery of unique QCs, in our previous study, we introduced a supervised learning workflow based on the XENONPY compositional descriptor by using the four summary statistics [27]. The input variable of the model was the chemical composition. The output variable represents its class label indicating three structural types: QCs, approximant crystals (ACs), and “others,” including ordinary periodic crystals. The chemical compositions of QCs, ACs, and ordinary crystals were used as training data. The trained predictive model described the three-class label as a function of the vectorized chemical composition. The kernel mean and summary statistics descriptors were compared on the predictive performances in this task. The procedure for generating both descriptors (such as the element features and setting of  $\sigma^2$  in the kernel mean descriptor) is the same as that used in the energy prediction task described above.

### C. Kernel mean force field descriptor on polymers

We constructed a kernel mean descriptor based on the empirical potential of all-atom classical molecular dynamics (MD) simulations to describe the chemical features of polymers. In the MD run, the motions of molecules are simulated according to the atomic interactions defined by the force-field potential. Various microscopic and macroscopic thermodynamic properties of the material system are calculated from the simulated trajectories. In principle, material properties are obtained owing to the nonlinear mapping of the potential function. Therefore, using the force-field parameters in the potential is natural as the description of material features.

In this paper, the General AMBER force field version 2 (GAFF2) [53] was employed to encode the chemical structure of a repeat unit in a linear polymer  $X$ . The potential energy is a function of the interatomic distance  $r = (r_{ij})_{ij}$  between atoms  $i$  and  $j$ , where all atom pairs  $(i, j)$  consisting of  $X$  are assigned to  $r$ . It is expressed as

$$\begin{aligned}
 U(r) = & \sum_{(i,j)} \epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \sum_{(i,j)} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}^2} \\
 & + \sum_{i \in \mathcal{N}_B} K_{\text{bond},i} (b_i - b_{0,i})^2 + \sum_{i \in \mathcal{N}_A} K_{\text{angle},i} (\theta_i - \theta_{0,i})^2 \\
 & + \sum_{i \in \mathcal{N}_D} K_{\text{dihedral},i} [1 + \cos(n_i \varphi_i - \varphi_{0,i})]. \quad (12)
 \end{aligned}$$

The first term, the vdW energy (Lennard-Jones potential), has two parameters describing the equilibrium distance  $\sigma_{ij}$  and the depth  $\epsilon_{ij}$  of the potential well. The second term is the potential energy, which defines the electrostatic interatomic interactions. The parameters consist of the charge  $q_i$  of each atom and dielectric constant  $\epsilon_0$  of the medium. These two potentials were defined for all pairs of atoms in the polymer. The third to fifth terms are the potential energies defining bond length stretching, bond angle expansion, and contraction, and dihedral angle rotation. These are defined for a set of atoms  $(\mathcal{N}_B, \mathcal{N}_A, \mathcal{N}_D)$  comprising the bond, angle, and dihedral angle. The equilibrium bond length  $b_{0,i}$ , force

TABLE I. List of 58 element features used to calculate the compositional descriptors. The data set is accessible with our open-source software XENONPY [48].

Feature ID	Description	Unit	Reference
atomic_number	Atomic number		
atomic_radius	Atomic radius	pm	[64,65]
atomic_radius_rahm	Atomic radius from Rahm <i>et al.</i>	pm	[65–67]
atomic_volume	Atomic volume	cm <sup>3</sup> mol <sup>-1</sup>	[65]
atomic_weight	Atomic weight		[65,68,69]
boiling_point	Boiling temperature	K	[65]
bulk_modulus	Bulk modulus	GPa	[70]
c6_gb	C <sub>6</sub> dispersion coefficient in a.u. (Gould & Bučko)	a.u.	[65,71–73]
covalent_radius_cordero	Covalent radius from Cordero <i>et al.</i>	pm	[65,74]
covalent_radius_pyykko	Single bond covalent radius from Pyykkö and Atsumi.	pm	[65,75]
covalent_radius_pyykko_double	Double bond covalent radius from Pyykkö and Atsumi.	pm	[65,76]
covalent_radius_pyykko_triple	Triple bond covalent radius from Pyykkö <i>et al.</i>	pm	[65,77]
covalent_radius_slater	Covalent radius from Slater	pm	[64]
density	Density at 295 K	g cm <sup>3</sup>	[65]
dipole_polarizability	Dipole polarizability	a.u.	[65,78]
electronegativity	Pauling electronegativity		[10]
electron_affinity	Electron affinity	eV	[65,79,80]
en_allen	Allen's scale of electronegativity	eV	[65,81,82]
en_ghosh	Ghosh's scale of electronegativity		[65,83]
en_pauling	Pauling's scale of electronegativity		[65,79]
first_ion_en	First ionization energy	eV	[79]
fusion_enthalpy	Enthalpy of fusion for elements at their melting temperatures	kJ mol <sup>-1</sup>	[79]
gs_bandgap	Density functional theory (DFT) band-gap energy at $T = 0$ K ground state	eV	[84,85]
gs_energy	DFT energy per atom (raw VASP value) at $T = 0$ K ground state	eV atom <sup>-1</sup>	[84,85]
gs_est_bcc_latcnt	Estimated BCC lattice parameter based on the DFT volume of the Open Quantum Materials database (OQMD) ground state for each element		[84,85]
gs_est_fcc_latcnt	Estimated FCC lattice parameter based on the DFT volume of the OQMD ground state for each element	Å	[84,85]
gs_mag_moment	DFT magnetic moment at $T = 0$ K ground state		[84,85]
gs_volume_per	DFT volume per atom at $T = 0$ K ground state	Å <sup>3</sup> atom <sup>-1</sup>	[84,85]
hhi_p	Herfindahl-Hirschman index (HHI) production values		[86]
hhi_r	HHI reserve values		[86]
heat_capacity_mass	Specific heat capacity at Standard temperature and pressure (STP)	J mol <sup>-1</sup> K <sup>-1</sup>	[79]
heat_capacity_molar	Molar heat capacity at STP	J mol <sup>-1</sup> K <sup>-1</sup>	[79]
icsd_volume	Volume per atom of Inorganic Crystal Structure Database (ICSD) phase at STP		[87–89]
evaporation_heat	Evaporation heat	kJ mol <sup>-1</sup>	[65]
heat_of_formation	Heat of formation	kJ mol <sup>-1</sup>	[65]
lattice_constant	Lattice constant	Å	[65]
mendelevov_number	Mendelevov's number		[65,90,91]
melting_point	Melting temperature	K	[65]
molar_volume	Molar volume	L mol <sup>-1</sup>	[70]
num_unfilled	Number of unfilled valence orbitals		[92,93]
num_valance	Number of valence electrons		[92,93]
num_d_unfilled	Number of unfilled <i>d</i> valence orbitals		[92,93]
num_d_valance	Number of filled <i>d</i> valence orbitals		[92,93]
num_f_unfilled	Number of unfilled <i>f</i> valence orbitals		[92,93]
num_f_valance	Number of filled <i>f</i> valence orbitals		[92,93]
num_p_unfilled	Number of unfilled <i>p</i> valence orbitals		[92,93]
num_p_valance	Number of filled <i>p</i> valence orbitals		[92,93]
num_s_unfilled	Number of unfilled <i>s</i> valence orbitals		[92,93]
num_s_valance	Number of filled <i>s</i> valence orbitals		[92,93]
period	Period in periodic table		[65]

TABLE I. (Continued.)

Feature ID	Description	Unit	Reference
specific_heat	Specific heat at 293.15 K	$\text{J g}^{-1} \text{mol}^{-1}$	[65]
thermal_conductivity	Thermal conductivity at 298.15 K	$\text{W m}^{-1} \text{K}^{-1}$	[65]
vdw_radius	van der Waals radius	pm	[65,79]
vdw_radius_alvarez	van der Waals radius from Alvarez	pm	[65,94]
vdw_radius_mm3	van der Waals radius from the MM3 force field. The MM3 force field is an advanced molecular mechanics (MM) force field which introduces the cross terms that involve up to three internal coordinates.	pm	[65,95]
vdw_radius_uff	van der Waals radius from the universal force field (UFF)	pm	[65,96]
sound_velocity	Velocity of sound	$\text{m s}^{-1}$	[70]
polarizability	Static average electric dipole polarizability	$10^{-24} \text{cm}^3$	[79]

constants of bond stretching  $K_{\text{bond},i}$ , equilibrium bond angle  $\theta_{0,i}$ , and force constants of bond bending  $K_{\text{angle},i}$  are assigned to each corresponding atom cluster. The last term is the potential for the dihedral angle, where  $K_{\text{dihedral},i}$ ,  $n_i$ , and  $\varphi_{0,i}$  are parameters. The force field descriptor was constructed with eight parameters with clear physical meanings. In addition, the atomic mass and bond polarity were added as parameters. The bond polarity is defined as the absolute value of the difference in charge between two atoms constituting a bond. Thus, the ten parameters listed in Table II were used in the descriptor set. The GAFF2 potential provides a predefined parameter set that is empirically determined for each element species or group [53].

Polymer  $X$  consists of components  $\{x_1, \dots, x_{N_x}\}$  and their relative frequencies  $\{w_1, \dots, w_{N_x}\}$ . Based on this, we calculated the summary statistics and kernel mean embedding for each of the ten force field parameters. If the number of grid points is set to  $d$  in the kernel mean descriptor, a descriptor vector of dimensions  $10 \times d$  is obtained. It should be noted here that, by definition, the support of a force field parameter

TABLE II. The ten types of parameters used in the force-field descriptor. The parameters are classified according to the assignment type: assignment for atom, bond, bond angle, and dihedral angle.

Assignment type	Parameter	Description
Atom	Mass	Atomic mass
	$\sigma$	Determining the equilibrium distance of vdW interactions
	$\epsilon$	Depth of the potential well of vdW interactions
	Charge	Atomic charge of the Gasteiger charge model
Bond	$r_0$	Equilibrium length of chemical bonds
	$K_{\text{bond}}$	Force constant of bond stretching
	Polar	Bond polarization defined by the absolute value of charge difference between atoms in a bond
Bond angle	$\theta_0$	Equilibrium angle of bond angles
	$K_{\text{angle}}$	Force constant of bond bending
Dihedral angle	$K_{\text{dihedral}}$	Rotation barrier height of dihedral angles

$\lambda$  is a discrete set, but the size of the set can be immense. For example, GAFF2 provides empirically determined values of the equilibrium bond lengths for 840 different element pairs. Therefore, designing a descriptor vector for all the data points that constitute the support considerably increases the dimensions of the vector. The dimensionality reduction using  $d$  discretized points is effective in reducing over-representation. In this study, the discretized points were ten points corresponding to ten different element species, such as hydrogen and carbon for mass, and 20 equally spaced grid points for the other parameters. Thus, the kernel mean force-field descriptor is given as a 190-dimensional vector.

In machine learning, the task is to build a predictive model that describes the thermophysical properties of linear polymers (e.g., thermal conductivity and coefficient of linear expansion) as functions of the vectorized chemical structure of any given polymer's repeating unit. Here, there is a technical problem to be overcome when dealing with the repeating units of a polymer in the descriptor calculation module. The repeating unit of a polymer cannot be determined uniquely. If a descriptor is calculated directly from the chemical structure of a given repeating unit, the substructure around the head-tail junction will not be included. Therefore, the descriptor was calculated after converting a repeating unit into a dimer or oligomer. For example, the SMILES notation for polyethylene is given as \*OCCC(=O)\* (the asterisks indicate the head and tail of the repeating unit), but by constructing the dimer \*OCCC(=O)OCCC(=O)\*, the structure around the junction C(=O)O can be reflected. However, the relative frequency of a substructure changed depending on the number of connected repeating units. To address this issue, we generated a SMILES string of a virtual polymer by infinitely repeating the given repeating unit structure by constructing a ten-unit oligomer and connecting its head and tail. The kernel mean and summary statistics descriptors were calculated using such virtually created SMILES strings. The descriptor calculations were performed using the RadonPy Python library [54].

#### IV. RESULTS

We demonstrate the predictive and expressive power of the kernel mean descriptor for the three different tasks described above.

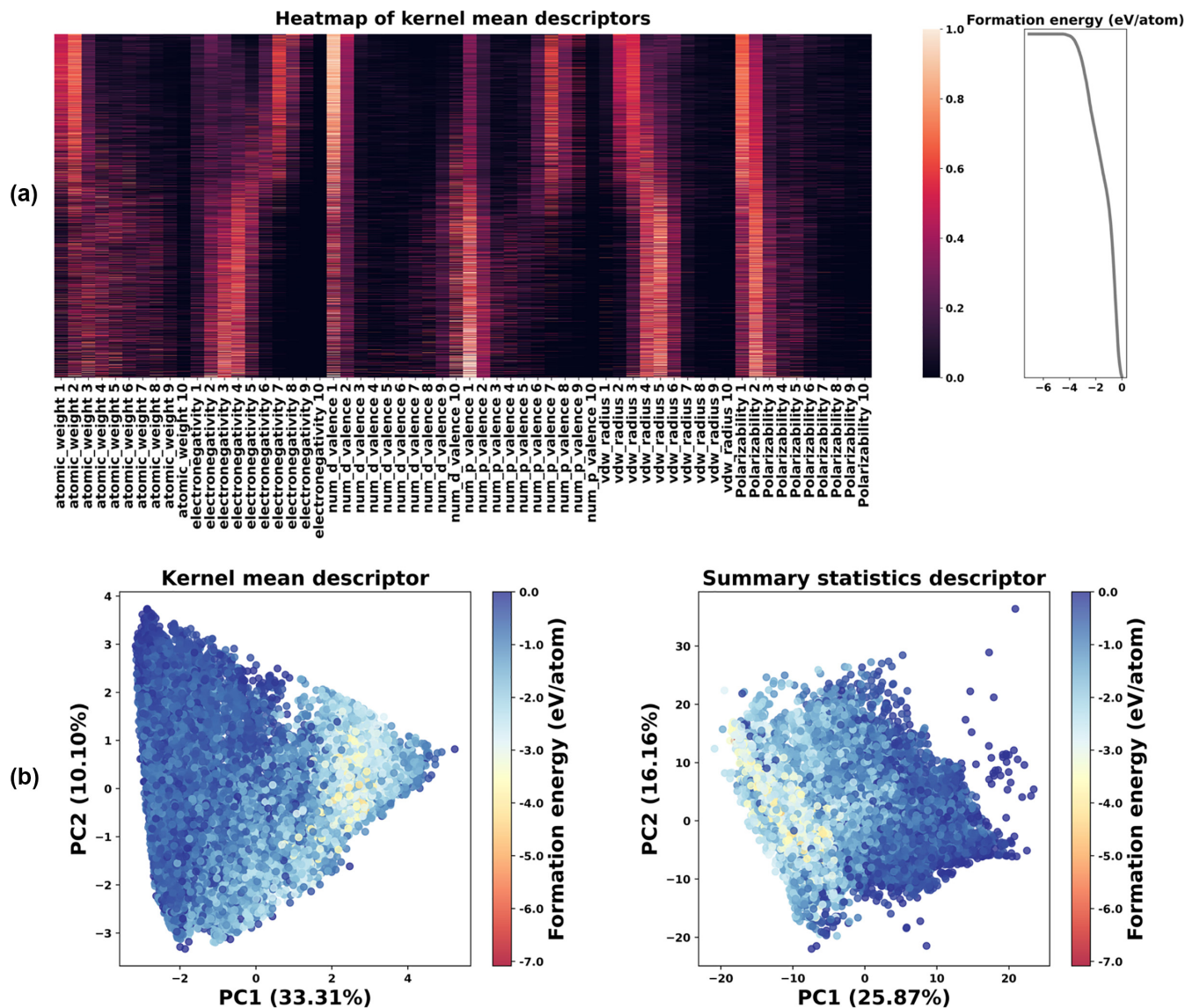


FIG. 2. Visualization of the kernel mean descriptors. (a) Heat map of the kernel mean descriptors of the chemical compositions for 35 463 stable inorganic compounds. The compounds were sorted by rows in the ascending order of formation energy (eV/atom). The kernel mean descriptors (shown on the columns) were generated with the ten equally spaced grid points ( $d = 10$ ) for six different element features (atomic weight, electronegativity, number of filled  $d$  valence orbitals, number of filled  $p$  valence orbitals, van der Waals radius, and polarizability). (b) Projection of the kernel mean descriptors (left) and the summary statistics descriptors (right) for the 35 463 compounds onto the first and second principal component axes (PC1 and PC2). The values in parentheses denote the cumulative contribution rates of PC1 and PC2. The colors vary from red (lower) to blue (higher) according to the magnitude of formation energies. The kernel mean descriptors were generated with 58 element features (see Table I for details) and ten equally spaced grid points ( $d = 10$ ), and the summary statistics descriptors were generated with the 58 element features and the four summary statistics as given by Eq. (2).

### A. Energy prediction of inorganic compounds

A total of 35 463 compounds were randomly divided into 21 277 ( $\sim 60\%$ ) for training, 7093 ( $\sim 20\%$ ) for validation, and 7093 ( $\sim 20\%$ ) for testing during the supervised learning process. As described in the Applications section, the 580-dimensional kernel mean descriptor and 232-dimensional summary statistics descriptor were generated to encode the compositional features. The mapping from a vectorized composition to the formation energy per atom was modeled and inferred using a conventional multilayer perceptron (MLP) with rectified linear units (ReLUs), fully connected lay-

ers, dropout layers, and batch normalization layers (see Appendix A for details of the model architecture and training procedure). By using the validation set, the optimized hyperparameters were selected from the following candidate solution sets by performing Bayesian optimization with the Optuna Python library [55]; the number of layers  $\in \{2, 3, 4\}$ , dropout rate  $\in \{0, 0.1, 0.2\}$ , and the number of neurons in each layer  $\in \{100, 150, 200, 250, 300\}$ . We then evaluated the generalization capability of the trained MLP on the test set. To account for uncertainty, the above procedure was repeated five times with different independently generated data splittings to



TABLE III. Performance of the formation energy prediction (eV/atom).

Descriptors	MAE	RMSE	R <sup>2</sup>
Kernel mean	0.0359 ( $\pm 0.0025$ )	0.0590 ( $\pm 0.0069$ )	0.9967 ( $\pm 0.0009$ )
Summary statistics	0.0413 ( $\pm 0.0006$ )	0.0658 ( $\pm 0.0070$ )	0.9959 ( $\pm 0.0009$ )

calculate the standard deviation of the performance metrics. The learning algorithm implemented in TensorFlow-macOS v2.9.0 was employed with the TensorFlow-metal v0.5.1 plugin for GPU calculations (Apple M1 Max, GPU 32 cores).

The mean absolute error (MAE), root-mean-square error (RMSE), and R<sup>2</sup> with respect to the test sets are presented in Table III. The performance metrics were averaged over the five trials, and the numbers in parentheses represent the standard deviations. It was confirmed that the kernel mean descriptor improves all three performance metrics relative to the summary statistics descriptor (from 0.0413 to 0.0359 eV/atom for MAE, 0.0658 to 0.0590 eV/atom for RMSE, and 0.9959 to 0.9967 for R<sup>2</sup>). Notably, the MAEs for the kernel mean and summary statistics descriptor reached 0.0359 eV/atom and 0.0413 eV/atom, respectively; these are less than the chemical accuracy of 1 kcal/mol (0.0434 eV/atom).

Figure 2(a) shows a heat-map display to visually understand the kernel mean descriptors of the 35 463 compounds in relation to the observed formation energies. For ease of visualization, only the six component features in the kernel mean descriptor are plotted ( $n(\Lambda) = 6$ ). Here, the compounds are arranged from top to bottom in the ascending order of formation energy. By visualizing the data set in this way, it was confirmed that a clear relationship exists between the distributional pattern of component features and the formation energy. As the formation energy increases, the electronegativity gradually changes from bimodal to unimodal. The larger the difference in electronegativity of the atoms in the crystal, the more strongly charged the atoms are. This increases the attraction of the Coulomb interaction and thus lowers the formation energy of the crystal. Therefore, the kernel mean descriptor represents the effect of the Coulomb interaction on the formation energy by expressing the multimodality of electronegativity. In addition, as the formation energy increases, the atomic radius increases (see `vdw_radius`). This can be interpreted as an increase in the formation energy because the Coulomb interaction decreases as the atomic radius increases.

Figure 2(b) shows the kernel mean descriptors (left) and summary statistic descriptors (right) of the 35 463 compounds projected onto the first and second principal component axes. The compounds were color-coded by the magnitude of formation energies on the two-dimensional coordinate axes of PCA to show the dependency between the compositional patterns and energy levels. The kernel mean descriptors are seen to be distributed inside the triangular region (left), and the formation energy is smoothly distributed within that region. In contrast, the summary statistics descriptors are generally distributed in the rectangular region (right); however, some compounds are scattered outside that region. In addition, the smoothness of the energy distribution has been lost in some areas, possibly causing a reduction in the prediction accuracy.

### B. Prediction of chemical composition to form quasicrystals

Following our previous work [27], we consider the problem of predicting QCs. The input variable of the model is a chemical composition. The output variable represents a class label indicating QC, AC, and others, including ordinary periodic crystals. As the data set, we used a list of the chemical compositions for 80 QCs and 78 ACs discovered to date [56]. For the others class, the chemical compositions of 10 000 periodic crystals were randomly extracted from the Materials Project database [50]. In addition, a list of 90 compositions was extracted from laboratory notebooks on failed syntheses of QCs that were added to the data set of the others class. A random forest classifier [57] was trained on 80% of the total data that was chosen at random; the number of training instances for QCs, ACs, and others were 64, 62, and 8072, respectively, and the remaining data were used for testing. By performing the fivefold cross validation within the training set, the hyperparameters were selected by Bayesian optimization using Optuna [55]; the parameters to be selected were the number of trees  $\in \{5, \dots, 1000\}$ , maximum depth of trees  $\in \{2, \dots, 100\}$ , number of features in each tree  $\in \{\text{sqrt}, \log_2\}$ , bootstrap sampling in the bagging  $\in \{\text{false}, \text{true}\}$ , and classification loss  $\in \{\text{entropy}, \text{Gini}\}$ . The predictive performance of the trained random forest classifier was evaluated on the test set. To quantify the reliability of the performance evaluation, we independently repeated the above procedure five times and calculated their standard deviation. The learning algorithm implemented in SCIKIT-LEARN [58] v1.1.3 was employed to train the models.

The recall, precision, F<sub>1</sub>, and macro F<sub>1</sub> on the test sets were calculated as summarized in Table IV. The classification task for the others class exhibited a significantly high recall, precision, and F<sub>1</sub> for both the kernel mean and summary statistics descriptors ( $>0.995$ ), indicating that the binary classification between the ordinary crystals (others) and the combined class of QCs and ACs is highly predictable based only on the compositional patterns of the already synthesized materials. In particular, the kernel mean descriptor exhibited better or equal predictive performance for all per-class performance metrics (recall, precision, and F<sub>1</sub>) relative to the summary statistic descriptor. Additionally, the kernel mean descriptor improved the macro F<sub>1</sub>, the mean of the per-class F<sub>1</sub> metrics, from 0.762 to 0.790 relative to the summary statistics descriptor.

### C. Kernel mean force-field descriptor on polymers

The chemical structure of a polymer repeating unit was used as the input to predict the thermal conductivity, linear expansion coefficient, and specific heat capacity at constant pressure ( $C_p$ ). The data set was generated by performing high-throughput all-atom MD simulations with five independent calculations for each of the 1138 linear polymers in amorphous states. Here, we used the LAMMPS MD software with

TABLE IV. Prediction performance of the kernel mean and summary statistics descriptors for the three-class classification task of stable QCs, ACs, and others. The table reports the per-class recall, precision,  $F_1$  metrics, and macro  $F_1$  that was calculated by taking the average of the per-class  $F_1$  over the three classes. The performance metrics were averaged over five independent trials, and the numbers in parentheses represent the standard deviations.

	Class	Recall	Precision	$F_1$	Macro $F_1$
Kernel mean	QC	0.562 ( $\pm 0.131$ )	0.798 ( $\pm 0.040$ )	0.653 ( $\pm 0.106$ )	0.790 ( $\pm 0.039$ )
	AC	0.662 ( $\pm 0.050$ )	0.791 ( $\pm 0.108$ )	0.718 ( $\pm 0.063$ )	
	Others	1.000 ( $\pm 0.000$ )	0.996 ( $\pm 0.001$ )	0.998 ( $\pm 0.001$ )	
Summary statistics	QC	0.538 ( $\pm 0.102$ )	0.765 ( $\pm 0.071$ )	0.629 ( $\pm 0.090$ )	0.762 ( $\pm 0.035$ )
	AC	0.612 ( $\pm 0.047$ )	0.727 ( $\pm 0.122$ )	0.661 ( $\pm 0.072$ )	
	Others	0.999 ( $\pm 0.001$ )	0.996 ( $\pm 0.001$ )	0.997 ( $\pm 0.001$ )	

the GAFF2 force field [59]. The computational details of the MD simulations are described in Ref. [59]. To obtain a reliable property data set, a set of polymers with at least three successful runs in repeated MD calculations was extracted from the entire data set, and the mean property value of each polymer was used as the observed output to be predicted. For thermal conductivity, samples with standard deviation  $> 0.05 \text{ W m}^{-1} \text{ K}^{-1}$  were excluded. Consequently, the number of samples was 996 for thermal conductivity, 1018 for linear expansion coefficient, and 1018 for  $C_p$ , respectively.

Out of the instances of the structural-property relationships, 80% of the randomly selected samples were used to train the Gaussian process (GP) regressor [60], and the remaining 20% were used as a test data set. The Matérn 5/2 kernel with automatic relevance determination was used for the kernel function in the GP. To adjust the kernel hyperparameters, the optimization was performed 50 times with different initial values to maximize the marginal likelihood.

The parity plots of the predicted and actual properties are shown in Fig. 3. Table V reports the mean and standard deviations of MAE, RMSE, and  $R^2$  for five independent experiments with different data splitting. For the prediction of thermal conductivity and linear expansion coefficient, the performance of the kernel mean descriptor was slightly better than that of the summary statistics descriptor. However, no improvement in the prediction performance of the kernel mean descriptor was observed for the prediction of  $C_p$  because both descriptors had almost perfect prediction performance ( $R^2 > 0.96$ ). The results suggest that the kernel mean descriptors improve the prediction performance for difficult prediction tasks in which the summary statistics descriptors are not sufficiently representative.

Here, we use the kernel mean descriptor to understand the input-output relationships embedded in the black-boxed

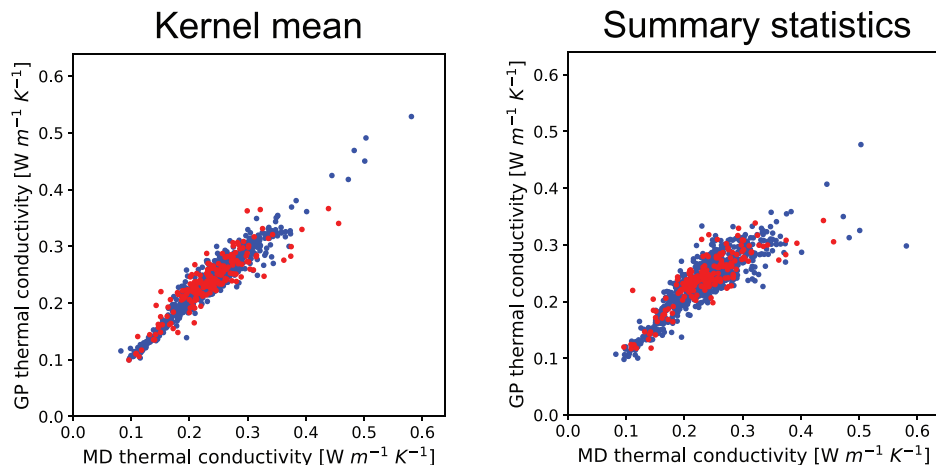
model. To evaluate the relevance of each feature in the kernel mean descriptor with respect to the thermal conductivity, we calculated the maximum information coefficient (MIC) [61], which is a measure of the strength of the linear or non-linear dependence between two random variables. Figure 4 shows the MIC score between each element of the kernel mean descriptor and the predicted thermal conductivity of the GP calculated for the 15 323 polymers recorded in PoLyInfo database [62]. In the bar plot of the MIC scores, the discretized regions for each force field parameter are arranged in ascending order from left to right.

The lower and higher ranges in the charge distribution were found to be highly relevant to the regulation of thermal conductivity, indicating that the proportion of atoms with a largely negative or positive charge is one of the dominant factors in determining thermal conductivity. In addition, a higher polar range was associated with a relatively higher MIC score; in other words, the proportion of highly polarized bonds is related to thermal conductivity. These results imply that strong electrostatic and dipole interactions are some of the controlling factors for the thermal conductivity of amorphous polymers. Moreover, a higher range of  $\epsilon$  had a high MIC score; thus, the proportion of atoms with strong vdW interactions is also highly relevant to thermal conductivity, suggesting that a strong vdW interaction is one of the factors controlling thermal conductivity. In our recent study, decomposition analyses of thermal conductivity based on physicochemical approaches concluded that strong nonbonding interactions improve the thermal conductivity of amorphous polymers [59]. In addition, a recent computational study on the contribution of localized vibrational modes to thermal conductivity in amorphous polymers showed that localized vibrational modes (locons) are the predominant mode types, and their contributions comprise more than 80% of

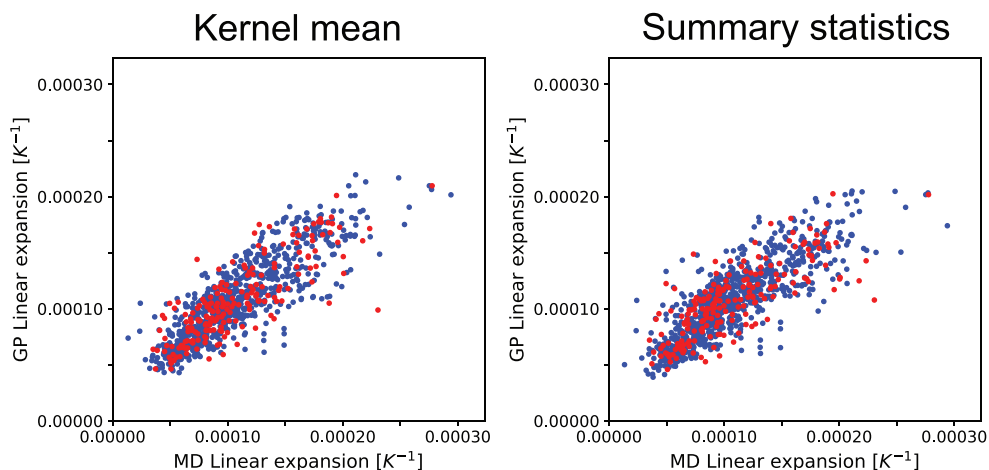
TABLE V. Prediction performance of the kernel mean and summary statistics descriptors with respect to three MD-calculated properties of linear polymers with amorphous states: thermal conductivity, linear expansion coefficient, and  $C_p$ .

Physical properties	Descriptors	MAE	RMSE	$R^2$
Thermal conductivity [ $\times 10^{-2} \text{ W m}^{-1} \text{ K}^{-1}$ ]	Kernel mean	2.15 ( $\pm 0.18$ )	3.21 ( $\pm 0.37$ )	0.677 ( $\pm 0.067$ )
	Summary statistics	2.40 ( $\pm 0.08$ )	3.29 ( $\pm 0.14$ )	0.662 ( $\pm 0.017$ )
Linear expansion coefficient [ $\times 10^{-5} \text{ K}^{-1}$ ]	Kernel mean	2.14 ( $\pm 0.20$ )	2.94 ( $\pm 0.26$ )	0.597 ( $\pm 0.052$ )
	Summary statistics	2.22 ( $\pm 0.13$ )	3.03 ( $\pm 0.21$ )	0.572 ( $\pm 0.040$ )
$C_p$ [ $\text{J kg}^{-1} \text{ K}^{-1}$ ]	Kernel mean	72.7 ( $\pm 7.6$ )	124.3 ( $\pm 19.8$ )	0.966 ( $\pm 0.011$ )
	Summary statistics	65.1 ( $\pm 6.7$ )	98.4 ( $\pm 10.9$ )	0.979 ( $\pm 0.004$ )

**(a) Thermal conductivity**



**(b) Linear expansion coefficient**



**(c)  $C_p$**

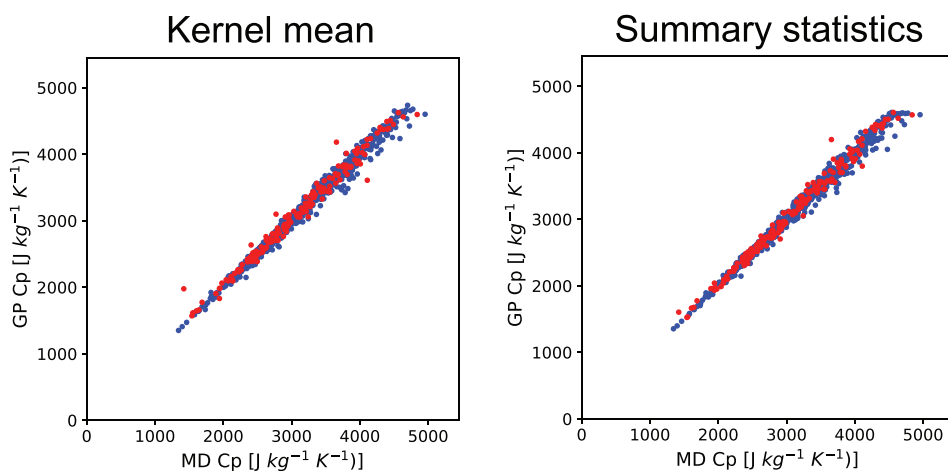


FIG. 3. Prediction results on (a) thermal conductivity, (b) linear expansion coefficient, and (c)  $C_p$  of linear polymers with amorphous states for the kernel mean embedding (left) and the summary statistics descriptor (right) of the force field parameters. Each parity plot shows GP-predicted properties against MD-calculated values. The training and test instances are color coded with blue and red, respectively.

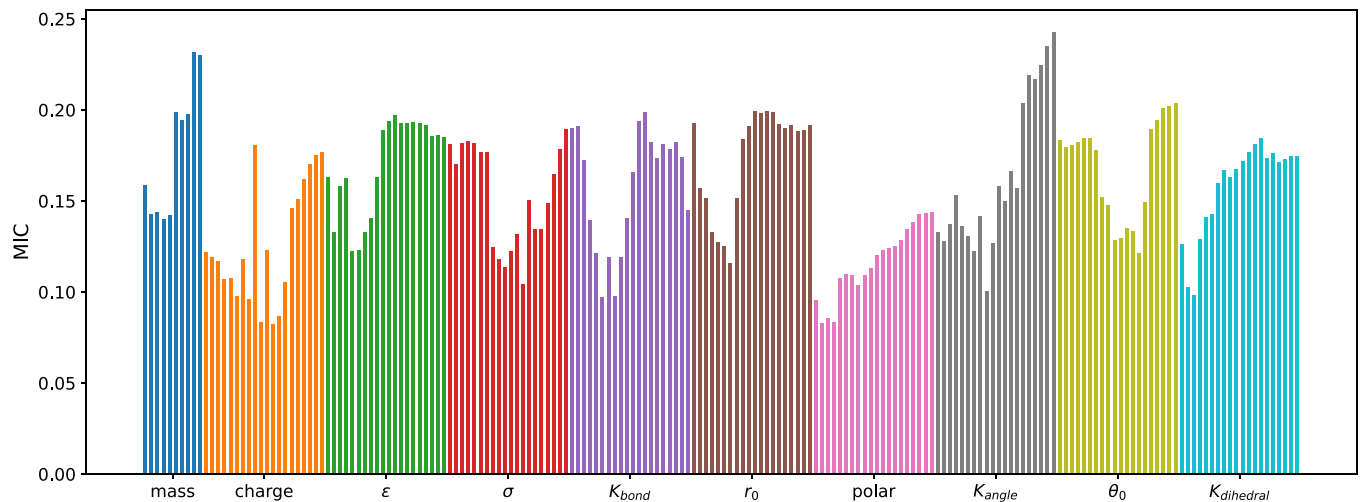


FIG. 4. MIC scores between ten force field parameters of the kernel mean force field descriptor and the predicted thermal conductivity. In the bar plot of the MIC scores, the discretized regions for each force-field parameter are arranged in ascending order from left to right.

the thermal conductivity [63]. More than half of the locon energy is confined to a single polymer chain, and minority locons are localized onto multiple polymer chains. This suggests that the latter modes play a key role in intermolecular heat transfer through vdW and electrostatic interactions. The observed MIC scores of charge, polarity, and  $\epsilon$  are consistent with the results of previous studies on their contributions to thermal conductivity in amorphous polymers. Therefore, the MIC results can be interpreted as strong nonbonding interactions that improve intermolecular heat transfer, which is the heat transfer bottleneck in amorphous polymers.

The higher and lower ranges of  $K_{\text{bond}}$  and a higher range of  $K_{\text{angle}}$  were highly related to thermal conductivity. These force-field parameters are related to the heat transfer in a single polymer chain. The decomposition analysis in our previous study indicated that the contribution of covalent bonds to thermal conductivity leads to the high thermal conductivity of polymers with rigid backbones, such as aromatic polyamides and polyimides [59]. In addition, as mentioned previously, locons localized in a single polymer chain contribute significantly to the thermal conductivity of amorphous polymers [63]. Therefore, we can interpret that these ranges of  $K_{\text{bond}}$  and  $K_{\text{angle}}$  are the controlling factors for heat transfer in a single polymer chain. It then follows as a natural hypothesis that because heat transfer in a single polymer chain has a more considerable contribution to the thermal conductivity of highly oriented polymers, further clarification of the mechanism by which  $K_{\text{bond}}$  and  $K_{\text{angle}}$  in the polymer chains affect heat transfer is important in the chemical design of oriented polymers with high thermal conductivity.

The same approach was applied to the linear expansion coefficient and  $C_p$  to reveal the input-output relationship of the black-box models. The results are described in Appendix B.

#### D. Predictive performances for varying descriptor dimensions

To investigate the effect of changes in the dimension of the kernel mean descriptor on prediction performances, we performed additional experiments on (a) the energy prediction of inorganic compounds and (b) the prediction of chemical

composition to form QCs, respectively. In tasks (a) and (b), prediction performances were evaluated on the kernel mean descriptors generated with  $d$  set to 3, 4, 7, 13, and 16, resulting in descriptor dimensions of 174, 232, 406, 754, and 928, respectively. All experimental conditions are the same as presented previously, except that the number of grid points  $d$  is changed. The change in prediction performance for varying the descriptor dimensions in tasks (a) and (b) is summarized in Fig. 5. The performance metrics were averaged over five independent trials, and their standard deviations are shown as error bars. The prediction performances at the 580-dimensional kernel mean descriptor ( $d = 10$ ) and the 232-dimensional summary statistics descriptor correspond to those reported in subsections A and B of the Results section.

As shown in Fig. 5(a), the mean MAE for the kernel average descriptor was best at 0.0357 eV/atom for the highest dimension of 928, and it was slightly better than that for a dimension of 580 (0.0359 eV/atom) (see also Fig. 6 for other performance metrics including RMSE and  $R^2$ ). In classification task (b), the mean macro  $F_1$  achieved the highest value of 0.790 with a dimension of 580. In task (a), the prediction performance tended to improve as the descriptor dimension increased, whereas in task (b) no such trend was observed. In task (a), where a sufficient amount of data was available (35 463 samples in total), increasing the descriptor dimension and improving the ability of the kernel mean descriptor to represent the distributional features would have been advantageous. The MAE changed significantly depending on the descriptor dimension, suggesting the importance of dimensionality selection in practice. Because the kernel mean descriptor is a discrete representation of the probability distribution function, the effective dimension does not change significantly as the number of grid points increases. This mechanism would suppress the performance degradation with increasing dimensionality.

The dimension of the summary statistics compositional descriptor is fixed at 232. This dimensionality corresponds to the case where  $d = 4$  in the kernel mean descriptor. Figure 5 shows a comparison of the prediction performance of two descriptors with the same dimension in tasks (a) and (b).



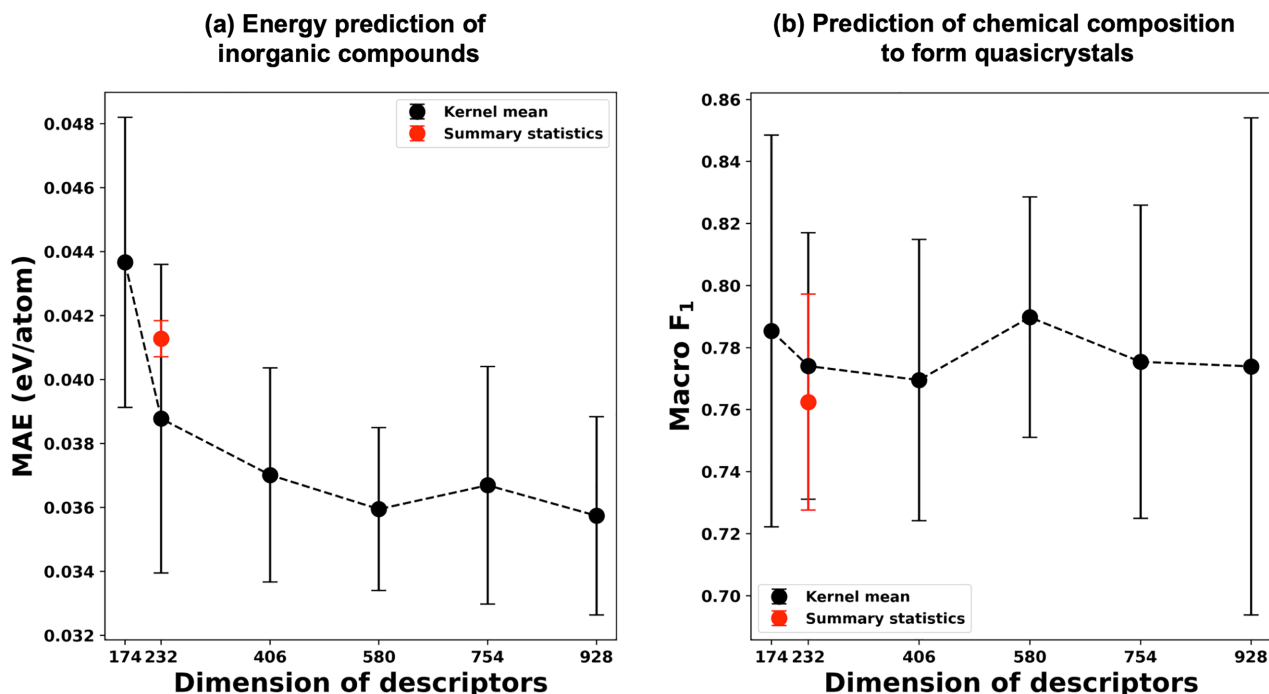


FIG. 5. Prediction performances for varying the dimension of the kernel mean descriptor on (a) the energy prediction task for inorganic compounds and (b) the task of predicting the chemical compositions that form QCs. For tasks (a) and (b), the black dots and error bars indicate the MAEs and the macro  $F_1$  averaged over five independent trials and their standard deviations, respectively. The red dots and their error bars at dimension 232 indicate the prediction performances of the summary statistics descriptor.

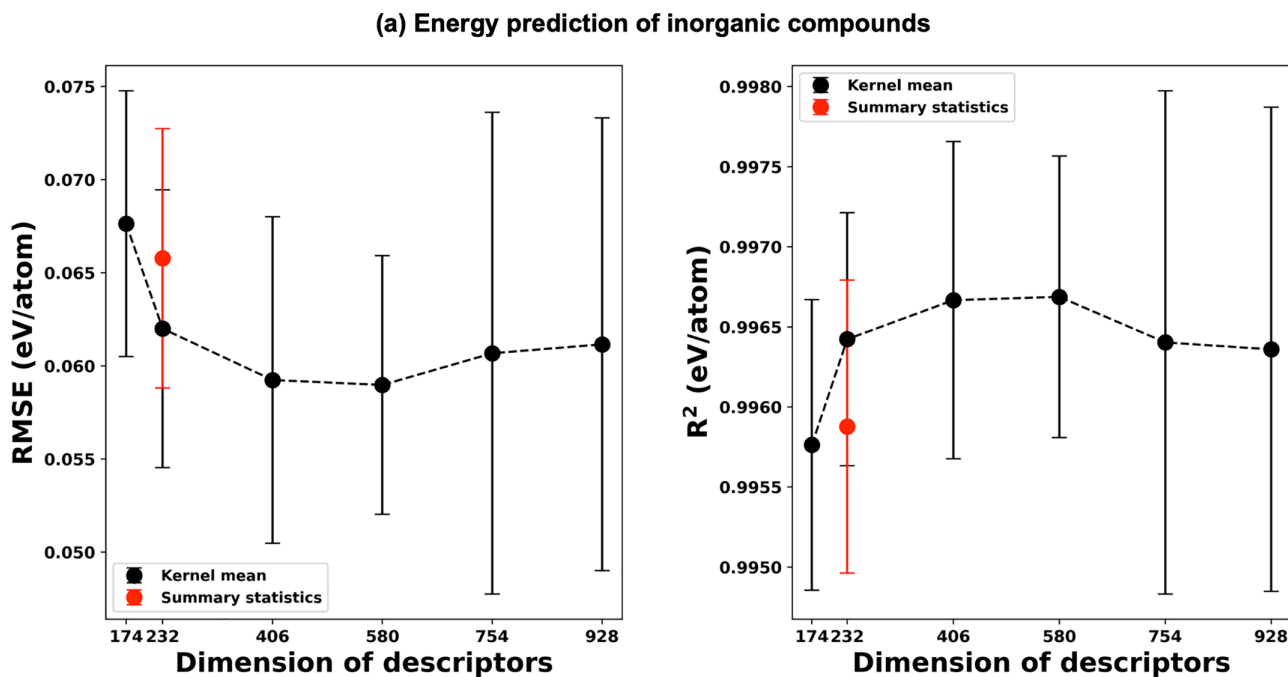


FIG. 6. Prediction performance curves for varying the dimension of the kernel mean descriptor on (a) energy prediction of inorganic compounds. The left and right figures use RMSE and  $R^2$  as performance metrics, respectively. For both figures, the black dots and error bars indicate the performance metric averaged over five independent trials and their standard deviations, respectively, for each dimension of the kernel mean descriptor. The red dot and error bar at dimension 232 indicates the performance metric averaged over five independent trials and their standard deviation, respectively, for the summary statistics descriptor.

In task (a), the mean MAE for the kernel mean descriptor with  $d = 4$  was 0.0388 eV/atom, which was much lower than 0.0413 eV/atom for the summary statistics descriptor. In task (b), the mean macro  $F_1$  for the kernel mean descriptor with  $d = 4$  was 0.774, which is higher than 0.762 for the summary statistics descriptor. This indicates that the kernel mean descriptor is able to compress compositional features into the same dimensional vector with high efficiency compared to the summary statistics descriptor.

## V. CONCLUSIONS

In materials research, the input material is often a mixed system consisting of multiple components, which can be expressed as a probability distribution. Therefore, material descriptors can be generated as vectorizing probability distributions. In this paper, we developed a general class of material descriptors based on the machine-learning theory of kernel mean embedding, which can map probability distributions to the feature space without losing any distributional features. We demonstrated the expressive power and versatility of the kernel mean descriptor in various applications, including the prediction of the formation energy of inorganic compounds, prediction of the chemical composition to form quasicrystalline materials, and use of force-field parameters to describe the compositional and structural features of polymer systems. Furthermore, by taking advantage of the linearity of the kernel mean embedding with respect to the component ratio, we presented an optimization framework that guarantees the uniqueness of the inverse transformation from the descriptor space to the material space.

A machine-learning property predictor defines a mapping  $Y = m(X)$  from a vector representation of material  $X$  to property  $Y$ . An ordinary descriptor defines a reduced representation of compositional and structural features of material  $X$ . Such descriptors treat several different materials as identical. Such overcollapsed representations unnecessarily constrain the expressive power of the resulting model. The kernel mean descriptor ensures that the mapping between  $X$  and  $\phi(X)$  is bijective whenever  $X$  is given as a mixture system defined by  $\{x_1, \dots, x_{N_x}\}$  and  $\{w_1, \dots, w_{N_x}\}$ . Once the complete distributional features of the input material are encoded in  $\phi(X)$ , a reduced representation should be obtained through machine learning of  $m$  in a data-driven manner. In particular, this notable feature of the kernel mean descriptor will bring an advantage when solving inverse problems, such as in high-throughput screening of unknown materials. Because overcollapsed descriptors can never discriminate between different materials with the same descriptor value, any model will recognize that their properties are identical, regardless of whether they exist inside or outside the training data distribution. This leads to the occurrence of many false positives and false negatives in the screening process, which can be avoided owing to the discriminative power of the kernel mean descriptor.

The kernel mean descriptor is operationally equivalent to the kernel density estimation. It is important to note the rationale for treating an inherently discrete distribution as a continuous distribution. For example, the possible values of a force field parameter are limited by the combination of

element species forming polymer systems. Thus, its support is a finite set. Similarly, for chemical composition descriptors, the domain of element features is limited by the number of existing element species. However, there is uncertainty in the given values of the force field parameters in an empirical potential as well as in the experimental and calculated values of the element features. To reflect the uncertainty of the predefined component features, they were represented based on continuous distributions.

The kernel mean descriptor has a wide range of potential applications, other than the three specific examples presented in this paper. In fact, the kernel mean descriptor can be applied to material systems of any dimension, such as 1D, 2D, and 3D, as long as there is a means to obtain a set of element-level features. In the geometrical representation of a crystalline system (a typical 3D scenario), the descriptor of the entire system is calculated by applying the averaging operation to component features  $\{\lambda_1, \dots, \lambda_{N_x}\}$  that represent the coordination environment of each atom ( $i \in \{1, \dots, N_x\}$ ). For example, in a crystal graph convolutional neural network, the component feature is obtained by repeatedly applying convolution operations to the feature vectors of each atom and its neighborhoods. Finally, crystal structures are encoded into fixed-length vectors by taking the sum of the component features. Here, this aggregation operation can be replaced by the kernel mean embedding. Molecular systems can also be treated as distributional objects whose constituent atoms and bonds differ in different systems. Furthermore, when searching the database for compositionally or structurally similar materials, a similarity measure that relies on the uniqueness of the kernel mean embedding will be used to better or perfectly discriminate between identical and nonidentical materials. Many other materials, including composite systems, copolymers, polymer blends, polymer solutions, and structural materials, can be represented within the unified framework of the kernel mean embedding.

The Python code for the kernel mean descriptor is available on GitHub [47]. The code can be generically used to create kernel mean descriptors for any mixture system. A program for the inverse translation of the kernel mean descriptors was also implemented in the code. The crystal data used in the formation energy prediction of inorganic compounds are available in the Materials Project, an open-access database [50,51]. A list of the QCs and ACs used in the prediction of the chemical composition to form the quasicrystalline phase is available in the Supporting Information of our previous work [27]. The Python code for the kernel mean force field descriptor was implemented in RADONPY, which is available on GitHub [54].

## ACKNOWLEDGMENTS

R.Y. acknowledges financial support from JST CREST Grants No. JPMJCR2203 and No. JPMJCR19I3, MEXT KAKENHI Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 19H50820), the Grant-in-Aid for Scientific Research (A) No. 19H01132 from the Japan Society for the Promotion of Science (JSPS), and the MEXT Program for Promoting Researches on the Supercomputer Fugaku (No. hp210264). Y.H. acknowledges financial support from a

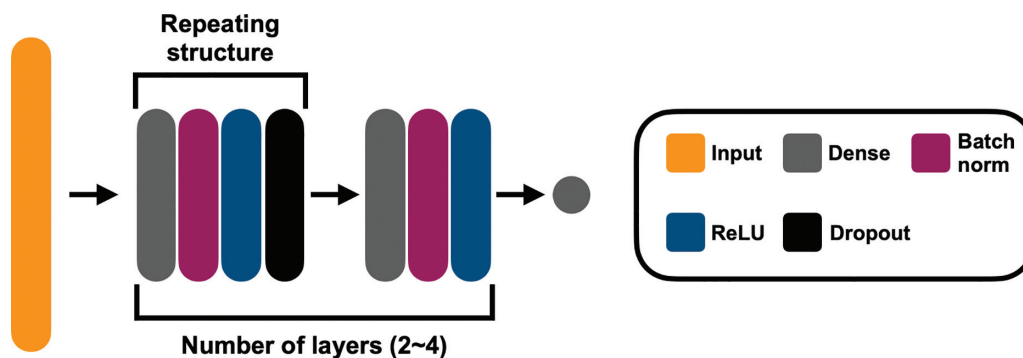


FIG. 7. Schematic view of the architecture of a conventional MLP model, which was employed for the formation energy prediction of inorganic compounds.

Grant-in-Aid for Scientific Research (C) No. 22K11949 from the Japan Society for the Promotion of Science (JSPS). The MD calculations for creating polymer properties were conducted using the Fugaku supercomputer at the RIKEN Center for Computational Science, Kobe, Japan, and the supercomputer at the Research Center for Computational Science, Okazaki, Japan (Project No. 22-IMS-C125).

M.K. and R.Y. designed the concept and outlined its proof. M.K. and R.Y. prepared the paper. M.K. conducted data analysis on the formation energy prediction of inorganic compounds and implemented a program for the inverse translation of kernel mean descriptors. M.K., A.W., and C.L. performed data analysis to predict the chemical composition to form the quasicrystalline phase. Y.H. wrote the paper, conducted the data analysis, and implemented the Python code for the force field descriptors for the polymer system. All authors discussed the results and commented on the paper.

The authors declare no conflicts of interest.

#### APPENDIX A: MLP ARCHITECTURE AND TRAINING

The architecture of a model employed for the formation energy prediction of inorganic compounds is shown in Fig. 7. This is a forward and fully connected neural network model, consisting of densely connected layers (dense), ReLUs, dropout layers (dropout), and batch normalization layers (batch norm). The network consists of one or more repeating structures with a dropout layer and an output layer with no dropout. All intermediate layers have the same number of units. As described in the Results section, by using the validation set, the hyperparameters of the models were selected by Bayesian optimization using the Optuna Python library (number of trials was set to 30) [55]. The solution space consists of the number of layers  $\in \{2, 3, 4\}$ , dropout rate  $\in \{0, 0.1, 0.2\}$ , and the number of neurons for each layer  $\in \{100, 150, 200, 250, 300\}$ . Each model was trained until the validation error converged (patience = 75 epochs) or until the number of epochs reached 1000. The validation error was evaluated as MAE. In each training process, we employed the model parameters that gave the lowest validation error throughout epochs as the final learned parameters. The Adam optimization technique [97] ( $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ) was used to back-propagate gradients. The batch size was fixed to 2048.

The learning algorithm implemented in TensorFlow-macOS v2.9.0 was employed to train the models with TensorFlow-metal v0.5.1 plug-in for GPU calculations (Apple M1 Max, GPU 32 cores).

#### APPENDIX B: INPUT—OUTPUT RELATIONSHIP OF THE FORCE FIELD DESCRIPTOR FOR LINEAR EXPANSION COEFFICIENT AND HEAT CAPACITY

To evaluate the relevance of each feature in the kernel mean force-field descriptor with respect to the linear expansion coefficient and  $C_p$  of polymers, we calculated the MIC score. Figure 8 shows the MIC score between each element of the kernel mean descriptor and the GP-predicted linear expansion coefficient and  $C_p$  for the 15 323 polymers recorded in PoLyInfo database [62].

For the linear expansion coefficient, higher MIC scores were observed in the high and low regions of both charge and polarity and in the low epsilon region [Fig. 8(a)]. Thus, the proportion of atoms with strong electrostatic and weak vdW interactions would play a significant role in determining the linear expansion coefficient. Electrostatic and vdW interactions are known to have highly anharmonic potential shapes; further, the thermal expansion of a material depends on the anharmonicity of the interatomic potentials. This suggests that the anharmonicity of the intermolecular potential is one of the controlling factors for the linear expansion coefficient. Significantly high MIC scores were also observed in the middle range of  $\theta_0$ . This range corresponds to a bond angle of  $105^\circ$ – $125^\circ$ , which is typical for  $sp^3$  or  $sp^2$  carbons. This indicates that the content of  $sp^3$  and  $sp^2$  carbons has a significant effect on the linear expansion coefficient. In general, polymers with more  $sp^2$  carbons tend to be more rigid because they have fewer rotatable bonds. The increase in rigidity decreases the micro-Brownian motion of the polymer chains, which in turn decreases the increase in free volume with increasing temperature.

As shown in Fig. 8(b), the pattern of the MIC in  $C_p$  is almost similar to that of the linear expansion coefficient; the MIC tended to be higher in the higher and lower ranges of charge and polar, the lower range of  $\epsilon$ , and the middle range of  $\theta_0$ . The similarity in MIC between  $C_p$  and the linear expansion coefficient suggests that the Grüneisen relation describing that thermal expansion coefficient is proportional to the heat

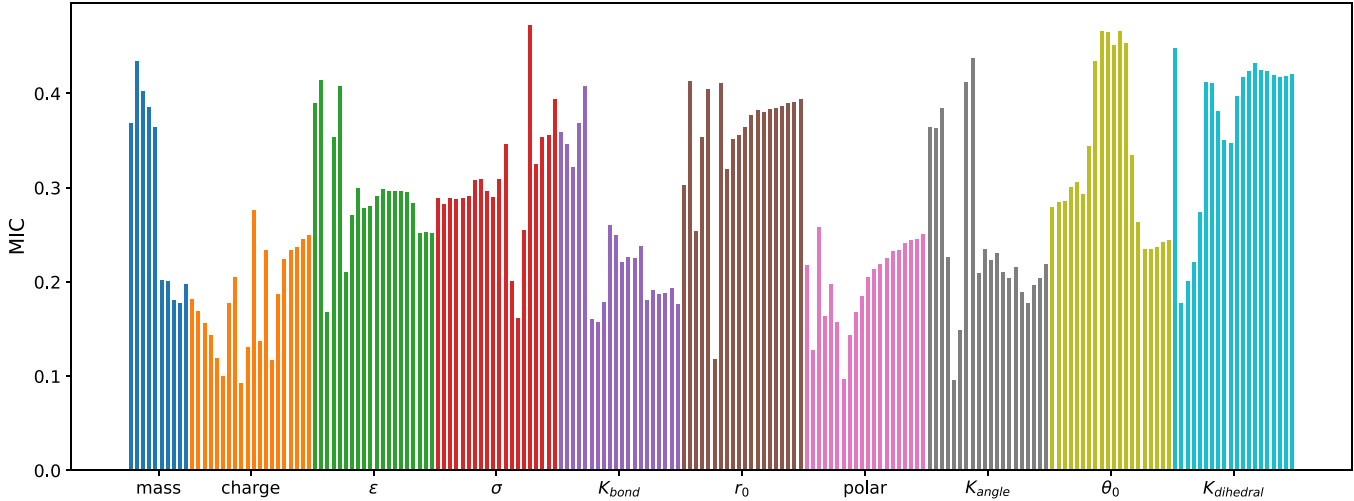
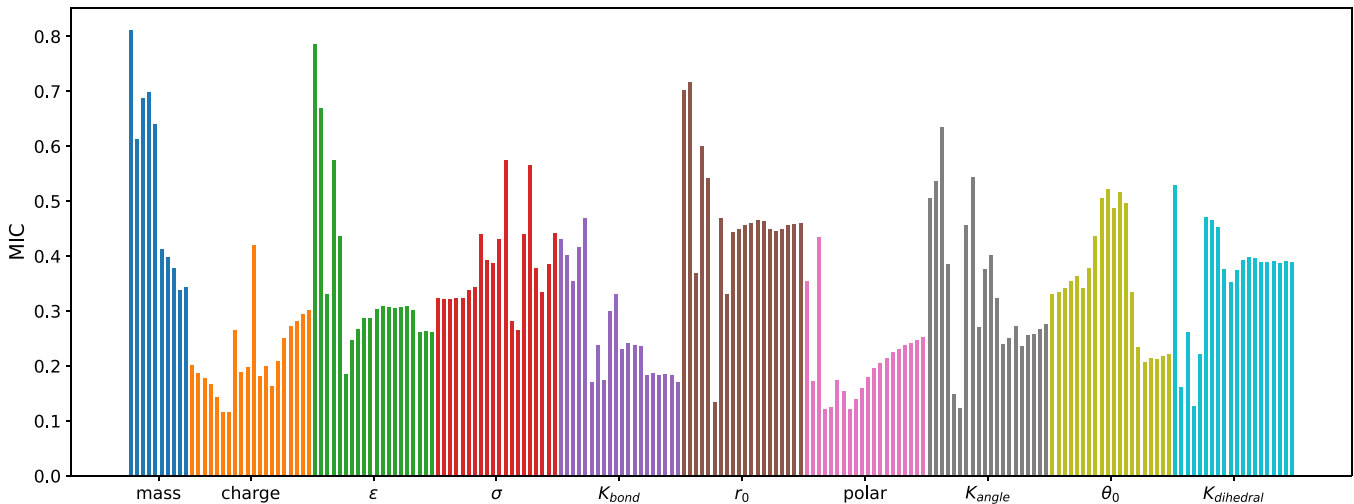
**(a) Linear expansion coefficient****(b)  $C_p$** 

FIG. 8. MIC scores between ten force field parameters of the kernel mean force field descriptor and the predicted linear expansion coefficient and  $C_p$  [in (a) and (b), respectively]. In the bar plot of the MIC scores, the discretized regions for each force-field parameter are arranged in ascending order from left to right.

capacity universally holds for amorphous polymers. In fact, our previous studies have shown that the MD-calculated

linear expansion coefficient and  $C_p$  are weakly proportional to each other [59].

- 
- [1] 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, and A. Aspuru-Guzik, *Nat. Mater.* **15**, 1120 (2016).
- [2] 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, and R. Yoshida, *npj Comput. Mater.* **5**, 66 (2019).
- [3] A. O. Oliyanyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, *Chem. Mater.* **28**, 7324 (2016).
- [4] R. Matsumoto, Z. Hou, H. Hara, S. Adachi, H. Takeya, T. Irifune, K. Terakura, and Y. Takano, *Appl. Phys. Express* **11**, 093101 (2018).
- [5] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, *Phys. Rev. Lett.* **115**, 205901 (2015).
- [6] J. Carrete, W. Li, N. Mingo, S. Wang, and S. Curtarolo, *Phys. Rev. X* **4**, 011019 (2014).
- [7] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, *Phys. Rev. B* **95**, 144110 (2017).
- [8] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).



- [9] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, *Nat. Commun.* **8**, 15679 (2017).
- [10] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, *npj Comput. Mater.* **2**, 16028 (2016).
- [11] D. Rogers and M. Hahn, *J. Chem. Inf. Model.* **50**, 742 (2010).
- [12] E. E. Bolton, Y. Wang, P. A. Thiessen, and S. H. Bryant, in *Annual Reports in Computational Chemistry* (Elsevier, 2008), Vol. 4, pp. 217–241.
- [13] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **42**, 1273 (2002).
- [14] R. E. Carhart, D. H. Smith, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **25**, 64 (1985).
- [15] R. Nilakantan, N. Bauman, J. S. Dixon, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **27**, 82 (1987).
- [16] K. Choudhary, B. DeCost, and F. Tavazza, *Phys. Rev. Mater.* **2**, 083801 (2018).
- [17] H. Ikebata, K. Hongo, T. Isomura, R. Maezono, and R. Yoshida, *J. Comput.-Aided Mol. Des.* **31**, 379 (2017).
- [18] S. Wu, G. Lambard, C. Liu, H. Yamada, and R. Yoshida, *Mol. Inf.* **39**, 1900107 (2020).
- [19] Y. Aoki, S. Wu, T. Tsurimoto, Y. Hayashi, S. Minami, O. Tadamichi, K. Shiratori, and R. Yoshida, *Macromolecules* **56**, 5446 (2023).
- [20] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, *J. Cheminf.* **10**, 4 (2018).
- [21] P. Broto, G. Moreau, and C. Vandycke, *Eur. J. Med. Chem.* **19**, 71 (1984).
- [22] G. Moreau and P. Broto, *Nouv. J. Chim.* **4**, 359 (1980).
- [23] P. A. Moran, *Biometrika* **37**, 17 (1950).
- [24] R. C. Geary, *The Incorporated Statistician* **5**, 115 (1954).
- [25] F. R. Burden, *J. Chem. Inf. Comput. Sci.* **29**, 225 (1989).
- [26] F. R. Burden, *Quant. Struct.-Act. Relat.* **16**, 309 (1997).
- [27] C. Liu, E. Fujita, Y. Katsura, Y. Inada, A. Ishikawa, R. Tamura, K. Kimura, and R. Yoshida, *Adv. Mater.* **33**, 2102507 (2021).
- [28] C. Liu, K. Kitahara, A. Ishikawa, T. Hiroto, A. Singh, E. Fujita, Y. Katsura, Y. Inada, R. Tamura, K. Kimura, and R. Yoshida, *Phys. Rev. Mater.* **7**, 093805 (2023).
- [29] A. R. Oganov and M. Valle, *J. Chem. Phys.* **130**, 104504 (2009).
- [30] N. E. Zimmermann and A. Jain, *RSC Advances* **10**, 6063 (2020).
- [31] T. Xie and J. C. Grossman, *Phys. Rev. Lett.* **120**, 145301 (2018).
- [32] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, *Advances in Neural Information Processing Systems*, Vol. 28 (Curran Associates, Inc., 2015).
- [33] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017).
- [34] Y.-J. Hu, G. Zhao, M. Zhang, B. Bin, T. Del Rose, Q. Zhao, Q. Zu, Y. Chen, X. Sun, M. de Jong *et al.*, *npj Comput. Mater.* **6**, 25 (2020).
- [35] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, *Found. Trends Mach. Learn.* **10**, 1 (2017).
- [36] A. Smola, A. Gretton, L. Song, and B. Schölkopf, in *International Conference on Algorithmic Learning Theory* (Springer, Berlin, Heidelberg, 2007), pp. 13–31.
- [37] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton, *Mach. Learn. Res.* **17**, 5272 (2016).
- [38] H. Xu, R. Liu, A. Choudhary, and W. Chen, *J. Mech. Des.* **137**, 051403 (2015).
- [39] G. E. Fasshauer, *Dolomites Res. Notes Approximation* **4**, 21 (2011).
- [40] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet, *J. Mach. Learn. Res.* **12**, 2389 (2011).
- [41] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, *Advances in Neural Information Processing Systems*, Vol. 20 (Curran Associates, Inc., 2007).
- [42] Y.-C. Chen, *Biostat. Epidemiol.* **1**, 161 (2017).
- [43] H. Takeda, S. Farsiu, and P. Milanfar, *IEEE Trans. Image Process.* **16**, 349 (2007).
- [44] W. S. Noble, *Nat. Biotechnol.* **24**, 1565 (2006).
- [45] quadprog: Quadratic programming solver (Python), <https://github.com/quadprog/quadprog>, accessed March 26, 2023.
- [46] D. Goldfarb and A. Idnani, *Math. Program.* **27**, 1 (1983).
- [47] Kmdplus, <https://github.com/Minoru938/KmdPlus>, accessed March 26, 2023.
- [48] Xenonpy platform, <https://github.com/yoshida-lab/XenonPy>, accessed March 26, 2023.
- [49] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, and R. Yoshida, *ACS Cent. Sci.* **5**, 1717 (2019).
- [50] The Materials Project, <https://materialsproject.org>, accessed Aug. 7, 2022.
- [51] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, *APL Mater.* **1**, 011002 (2013).
- [52] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, *Phys. Rev. Lett.* **53**, 1951 (1984).
- [53] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, *J. Comput. Chem.* **25**, 1157 (2004).
- [54] Radonpy GitHub site, <https://github.com/RadonPy/RadonPy>, accessed March 26, 2023.
- [55] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, Anchorage AK USA, 2019), pp. 2623–2631.
- [56] S. Walter and S. Deloudi, *Crystallography of Quasicrystals: Concepts, Methods and Structures*, Springer Series in Materials Science, Vol. 126 (Springer, Berlin, 2009), pp. 261–271.
- [57] T. K. Ho, in *Proceedings of 3rd International Conference on Document Analysis and Recognition* (IEEE, Montreal, QC, Canada, 1995), Vol. 1, pp. 278–282.
- [58] 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, and E. Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [59] Y. Hayashi, J. Shiomi, J. Morikawa, and R. Yoshida, *npj Comput. Mater.* **8**, 222 (2022).
- [60] C. E. Rasmussen, *Gaussian Processes in Machine Learning* (Springer, Berlin, Heidelberg, 2003), pp. 63–71.
- [61] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, *Science* **334**, 1518 (2011).
- [62] S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu, and M. Yamazaki, in *2011 International Conference on Emerging Intelligent Data and Web Technologies* (IEEE, Tirana, Albania, 2011), Vol. 22.

- [63] B. Li, F. DeAngelis, G. Chen, and A. Henry, *Commun. Phys.* **5**, 323 (2022).
- [64] J. C. Slater, *J. Chem. Phys.* **41**, 3199 (1964).
- [65] Mendeleev—a python resource for properties of chemical elements, ions and isotopes, ver. 0.3.6, <https://github.com/lmmementel/mendeleev>, accessed March 26, 2023.
- [66] M. Rahm, R. Hoffmann, and N. W. Ashcroft, *Chem. Eur. J.* **22**, 14625 (2016).
- [67] M. Rahm, R. Hoffmann, and N. W. Ashcroft, *Chem. Eur. J.* **23**, 4017 (2017).
- [68] J. Vogt and S. Alvarez, *Inorg. Chem.* **53**, 9260 (2014).
- [69] J. Meija, T. B. Coplen, M. Berglund, W. A. Brand, P. De Bièvre, M. Gröning, N. E. Holden, J. Irrgeher, R. D. Loss, T. Walczyk, and T. Prohaska, *Pure Appl. Chem.* **88**, 265 (2016).
- [70] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
- [71] K. T. Tang, J. M. Norbeck, and P. R. Certain, *J. Chem. Phys.* **64**, 3063 (1976).
- [72] X. Chu and A. Dalgarno, *J. Chem. Phys.* **121**, 4083 (2004).
- [73] T. Gould and T. Bučko, *J. Chem. Theory Computation* **12**, 3603 (2016).
- [74] B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán, and S. Alvarez, *Dalton Trans.*, 2832 (2008).
- [75] P. Pyykkö and M. Atsumi, *Chem. Eur. J.* **15**, 12770 (2009).
- [76] P. Pyykkö and M. Atsumi, *Chem. Eur. J.* **15**, 186 (2009).
- [77] P. Pyykkö, S. Riedel, and M. Patzschke, *Chem. Eur. J.* **11**, 3511 (2005).
- [78] P. Schwerdtfeger and J. K. Nagle, *Mol. Phys.* **117**, 1200 (2019).
- [79] W. M. Haynes, *CRC Handbook of Chemistry and Physics*, 95th ed. (CRC Press, Oakville, 2014).
- [80] T. Andersen, *Phys. Rep.* **394**, 157 (2004).
- [81] J. B. Mann, T. L. Meek, and L. C. Allen, *J. Am. Chem. Soc.* **122**, 2780 (2000).
- [82] J. B. Mann, T. L. Meek, E. T. Knight, J. F. Capitani, and L. C. Allen, *J. Am. Chem. Soc.* **122**, 5132 (2000).
- [83] D. C. Ghosh, *J. Theor. Comput. Chem.* **04**, 21 (2005).
- [84] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, *JOM* **65**, 1501 (2013).
- [85] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, *npj Comput. Mater.* **1**, 15010 (2015).
- [86] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, *Chem. Mater.* **25**, 2911 (2013).
- [87] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, *Acta Crystallogr. Sect. B Struct. Sci.* **58**, 364 (2002).
- [88] R. Allmann and R. Hinek, *Acta Crystallogr. Sect. A Found. Crystallogr.* **63**, 412 (2007).
- [89] D. Zagorac, H. Muller, S. Ruehl, J. Zagorac, and S. Rehme, *J. Appl. Crystallogr.* **52**, 918 (2019).
- [90] D. G. Pettifor, *Solid State Commun.* **51**, 31 (1984).
- [91] P. Villars, K. Cenzual, J. Daams, Y. Chen, and S. Iwata, *J. Alloys Compd.* **367**, 167 (2004).
- [92] Material-agnostic platform for informatics and exploration, <https://wolverton.bitbucket.io>, accessed May 26, 2023.
- [93] Electron configuration of the elements, <https://periodictable.com/Properties/A/ElectronConfigurationString.v.html>, accessed May 26, 2023.
- [94] S. Alvarez, *Dalton Trans.* **42**, 8617 (2013).
- [95] N. L. Allinger, X. Zhou, and J. Bergsma, *J. Mol. Struct. (THEOCHEM)* **312**, 69 (1994).
- [96] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, *J. Am. Chem. Soc.* **114**, 10024 (1992).
- [97] D. P. Kingma and J. Ba, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).