A general representation scheme for crystalline solids based on Voronoi-tessellation real feature values and atomic property data

Randy Jalem\textsuperscript{a,}\textsuperscript{b,} Masanobu Nakayama\textsuperscript{b,}\textsuperscript{c,}\textsuperscript{d,}\textsuperscript{e} Yusuke Noda\textsuperscript{a}, Tam Le\textsuperscript{f}, Ichiro Takeuchi\textsuperscript{i,}\textsuperscript{j}, Yoshitaka Tateyama\textsuperscript{b,}\textsuperscript{e} and Hisatsugu Yamazaki\textsuperscript{h}

\textsuperscript{a}Japan Science and Technology agency (JST), PRESTO, Saitama, Japan; 
\textsuperscript{b}National Institute for Materials Science – Global Research Center for Environment and Energy Based on Nanomaterials Science (NIMS – GREEN), Tsukuba, Japan; 
\textsuperscript{c}Center for Materials Research by Information Integration (cMi2), Research and Services Division of Materials Data and Integrated System (MaDIS), National Institute for Materials Science – ‘Materials Research by Information Integration’ Initiative (NIMS – Mi2i), Tsukuba, Japan; 
\textsuperscript{d}Frontier Research Institute for Materials Science, Nagoya Institute of Technology, Nagoya, Japan; 
\textsuperscript{e}Elements Strategy Initiative for Catalysts and Batteries, Kyoto University, Kyoto, Japan; 
\textsuperscript{f}Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan; 
\textsuperscript{g}RIKEN Center for Advanced Intelligence Project, Tokyo, Japan; 
\textsuperscript{h}Battery Material Engineering & Research Division, Toyota Motor Corporation, Susono, Japan

ABSTRACT

Increasing attention has been paid to materials informatics approaches that promise efficient and fast discovery and optimization of functional inorganic materials. Technical breakthrough is urgently requested to advance this field and efforts have been made in the development of materials descriptors to encode or represent characteristics of crystalline solids, such as chemical composition, crystal structure, electronic structure, etc. We propose a general representation scheme for crystalline solids that lifts restrictions on atom ordering, cell periodicity, and system cell size based on structural descriptors of directly binned Voronoi-tessellation real feature values and atomic/chemical descriptors based on the electronegativity of elements in the crystal. Comparison was made vs. radial distribution function (RDF) feature vector, in terms of predictive accuracy on density functional theory (DFT) material properties: cohesive energy (CE), density (d), electronic band gap (BG), and decomposition energy (Ed). It was confirmed that the proposed feature vector from Voronoi real value binning generally outperforms the RDF-based one for the prediction of aforementioned properties. Together with electronegativity-based features, Voronoi-tessellation features from a given crystal structure that are derived from second-nearest neighbor information contribute significantly towards prediction.

1. Introduction

Materials informatics in tandem with high-throughput density functional theory (DFT) calculations has now become increasingly deployed for exploration, optimization, and property forecasting of inorganic materials. This became evident with the emergence of DFT databases containing a large number of entries on physical and thermodynamical properties, for example,
for data mining and material screening tasks [1–3]. Aside from the wealth of material data-sets, the inorganic materials field has also greatly benefited from advances made in machine learning algorithms and computer hardware performance. Practical demonstrations and successes have since been shown in areas such as Li-ion batteries [4–9], photovoltaic devices [10], catalysis [11,12], thermoelectrics [13,14], solar cells [15], superconductors [16], and data storage [17].

The underlying assumption in the use of informatics tools that affords learning of structure–property relationships is the effectiveness and validity of the input data representation for materials, that is, descriptors [18]. The choice and formulation for this is often far from being trivial but in common practice, it is usually guided by prior knowledge of a given application domain or by chemical intuition. For example, we used local structure descriptors and atomic/chemical descriptors in our previous works in order to efficiently predict a transition state property called Li ion migration energy which is a useful metric for assessing fast ionic conduction in candidate compounds for use in battery applications [8,18–20]. Although the structure descriptors were proven predictive for the aforementioned target property, most of them were inherently constrained and usable only to a few structure types at a time. For more general chemical search spaces or machine learning tasks though, structure descriptors need to be unrestricted with respect to structure type. Meanwhile, it is worth mentioning that atomic or chemical descriptors have been used as well by others with notable success for a variety of materials property predictions [18,21,22].

In the bio- and chem-informatics field, chemical graph theoretical matrices are commonly applied to represent molecules in a general manner [23]. Not surprisingly, some of the early representations for inorganic materials were inspired from efforts such as this, example is the so-called Coulomb matrix (CM) which contains information on nuclear charge $Z$ and atom $I$-atom $J$ pair distance $r$ [24,25]:

$$m_{I,J} = \begin{cases} 0.5Z_I^2 \frac{1}{r^{I,J}} & \text{for } I = J \\ \frac{Z_I Z_J}{|r^{I,J}|} & \text{for } I \neq J \end{cases},$$

where $m_{I,J}$ denotes matrix element. Essentially, the method is a form of descriptor binning with the matrix elements $m_{I,J}$ assigned for atom-atom pair bins. However, the direct use of input representations from organic molecules for inorganic materials is usually inappropriate and ineffective because of 3D periodicity issue and dependency on the number of atoms and atom ordering. CM was combined with Bravais matrix (i.e. information from primitive translation vectors and unit cell basis) as a feature vector but the predictive accuracy is unsatisfactory, e.g. for the value of the density of states at the Fermi energy (DOS$_F$) [26].

Alternative representations for inorganic crystals have since been proposed, again, within the concept of binning material features. One example is the concatenation of a series of radial distribution functions (RDFs) of various atom type pairs:

$$g(r) = 4\pi r^2 \rho dr,$$

where $\rho$ represents number density of the unit cell, to directly and indirectly embed geometric and chemical information, respectively, into a vector or list [8,26,27]. RDF-based feature vector addresses cell periodicity issue by excluding interatomic distance contributions beyond a certain cutoff value. However, when all element combinations (~5000) are considered, RDF descriptors can easily become enormously large in dimensionality and be padded in large portion with zeros that do not contribute to learning. One recently reported representation shows good promise for solids and molecules and uses a kernel function $K$ to compare, say between two solids $A$ and $B$, based on smooth overlap of atomic positions [28]:

$$K(A, B) = \sum_{i\in A, j\in B} P_{ij} k(x_i, x_j),$$

where $x_i, x_j$ relates to atom-centered environments and $P_{ij} = \frac{1}{n^{XX}}$. Another representation attempted to include electronic structure information by binning energy diagrams of high symmetry points of a crystal’s Brillouin zone and binning of DOS data [16]. Truncated bispectrum approach was introduced to generate descriptors based on projected atomic density information onto a 4-D unit sphere surface for one-to-one representation of atomic neighborhoods [29]:

$$B_{J_1J_2J_3} = \sum_{m_1, m_2, m_3} \sum_{m_1' j_1 m_2' j_2 m_3' j_3} \left(c_{m_1' m_2' m_3' j_1 j_2 j_3} \right)^* C_{J_1 m_1 j_1}^{m_1} C_{J_2 m_2 j_2}^{m_2} C_{J_3 m_3 j_3}^{m_3},$$

where $C_{Jm}$ are 4D spherical harmonics coefficients related to atom index $J$ (atom index $I$ is dropped for clarity in the equation), $C_{Jm}^{m'}$ are ordinary Clebsch-Gordan coefficients, and $J_1 J_2 J_3$ denote the spatial resolution limit defining the atomic neighborhood.

Another class of material representation was defined by decomposing infinite and periodic crystal structures into finite number of representative fragments based on rules of chemical bonding coordination sphere determined from Voronoi-tessellation cells of atom centers. The shape of a Voronoi cell is often described by three integer sets: $[F_i] [V_i] [E_i]$ where $F$, $V$, $E$ are the number of $i$-edged faces, number of vertices, and number of edges, respectively, and $i \in \{3, 4, 5, \ldots \}$ [30,31]. In one study, Voronoi cell information was combined with atomic/chemical labeling in order to define crystal substructures for applications in structure similarity metrics and crystal site prediction [32]. In yet another study, a feature
vector was formulated with a similar idea, constructing at first infinite 3D periodic graphs of vertices \( V \) (atoms) and edges \( E \) (bonding):

\[
G = (V, E),
\]

and then eliminating the problem on cell periodicity by breaking down the conceptually infinite graphs into unit-cell-relevant subgraph types from which finite-sized adjacency matrices were then determined:

\[
A_{ij} = \begin{cases} 
1 & \text{if } I \text{ and } J \text{ are adjacent} \\
0 & \text{otherwise}
\end{cases},
\]

where \( A \) is for adjacency matrix element encompassing all atom pairing \( I \) and \( J \) in one subgraph type [33]. In these schemes, it is obvious that: (i) both chemical/atomic and geometrical properties are always considered in an attempt to make the representation universal, (ii) atomic or chemical descriptors from the literature may be made suitable for defining the ‘chemical identity’ of materials, and (iii) Voronoi cell features may provide the criteria to describe the ‘structure identity’ of materials.

In this work, we further explored the use of Voronoi tessellation for differentiating crystal structures of inorganic materials. However, instead of utilizing Voronoi features as criteria for formulating crystal substructures or fragments for defining ‘structure identity’, we directly binned the Voronoi feature real values themselves and utilize the bin count information to construct general vector-form descriptors. For ‘chemical identity’, histogram from atomic/chemical property data was added into the overall material fingerprint. The approach was validated by performing supervised learning on DFT-calculated properties for inorganic materials: cohesive energy (CE), density (\( d \)), electronic band gap (BG), and decomposition energy (Ed). Our proposed scheme offers a simple, robust and cost-effective alternative for generating feature vector representations for a wide variety of crystalline materials.

2. Computational details

2.1. DFT calculation, material data-set, target properties

The material properties that were used for supervised learning are CE, \( d \), BG, and Ed. For generating the material data-set, DFT calculations were performed with the VASP code which uses projected augmented wave (PAW) potentials [34,35]. The standard generalized gradient approximation (GGA) with Perdew–Burke–Ernzerhof parametrization for solids (PBEsol) was applied to describe exchange correlation effects [36]. Under spin-polarized condition, structure geometry optimization was carried out with a cutoff of 500 eV in kinetic energy and this led to a convergence upper bound of 1 meV per formula unit in the total energy. K-point mesh was set to \( \geq 1000 \) and with Monkhorst-Pack grids.

We focused our search space to Li-containing compounds owing to their importance as fast ionic conductors for batteries. Structure coordinate data were collected from the Inorganic Crystal Structure Database (ICSD) [37]. Separate 1000 data-sets were taken as well from Materials Project [1], for analyzing the effect of training data-set size on prediction performance. Transition metal elements were excluded to generate a rather homogeneous data-set avoiding entries that may contribute large uncertainties in error measures related BG calculation (usually underestimated by DFT). Duplicates and unphysical structures were removed to form the model building and test sets. Additional information is available in Table S1 (ICSD-based DFT data) and S2 (material ID of Materials Project data) of Supporting Information.

To determine CE, the energy of atoms in the crystal and the energy of corresponding free atoms need to be calculated from DFT:

\[ CE = E_{\text{crystal}} - \sum_{i} n_i E_i, \]

where \( E_{\text{crystal}} \) is the total energy of the compound, \( N_{\text{atom types}} \) is the number of atom types, \( n_i \) is the number of atoms of type \( i \), and \( E_i \) is the total energy of an atom type \( i \) (calculated by placing a single atom in the middle of a \( 20 \times 20 \times 20 \) Å simulation box. Standard DFT typically underestimates BG by as much as 50%, particularly for insulators and semiconductors [38]. This discrepancy is usually solved by introducing the Hubbard \( U \) extension [39,40] or the GW many-body scheme [41,42] but the computation is extremely expensive, especially with respect to system size. Also, calculation for direct BG requires high k-point sampling and stricter energy convergence criterion. In this work, we considered the energy difference between the valence band top and conduction band bottom for DFT-BG estimation.

2.2. Formulation of generalized vector-form material descriptors

We took inspiration from the idea of binning and the frequency of occurrence for material-related features to generate ‘histograms of features’. We concatenated the histograms from atomic and geometric information around atom centers and use it to define a generalized input vector representation for any given material \( x \):

\[
x = [h_{i}]^{k_i} \overset{c}{\sim} [h_{j}]^{k_j},
\]

where \( [\cdot]_c \) and \( [\cdot]_j \) are histograms for the ‘chemical’ and ‘structure’ identities of the material, respectively, joined together by a concatenation operator \( \overset{c}{\sim} \), \( h_i \) is the atomic
feature count at bin $i$ where $i \in [1, 2, \ldots, k_1]$ ($k_1$ total bins), $h_j$ is the local geometric feature count at bin $j$ where $j \in [1, 2, \ldots, k_2]$ ($k_2$ total bins).

For $[\cdot]$, we chose the atomic property based on Pauling's electronegativity of elements (EN) [43]. The reason for this choice is that it allows for defining chemical bonding in crystalline solids on a continuum or scale, whereby bonds can be mapped as strongly ionic on one extreme, strongly covalent on the other, or somewhere in between (i.e. a combination of both ionic and covalent character). Other atomic-related properties can also be considered and added as well such as electron affinity, atomization enthalpies, vaporization enthalpies, fusion enthalpies, melting point, atomic mass, atomic number, number of valence electrons, ionization potentials, covalent radii, ionic radii [44], and element periodic group number. However, it should be pointed out that further increasing the dimensionality of the material representation would also require an appropriate increase in the size of training set so as to reduce the risk of optimistic and biased evaluations on machine model performance [45]. In the present work, we ensure to mitigate the latter by performing subsampling (of training set and feature set), regularization, and cross-validation. Meanwhile, $[\cdot]$, was derived from Voronoi-based numerical feature values around each atoms in a crystal structure [46].

Figure 1 illustrates the proposed concept for generating atomic- and geometric-based binned features. Several local Voronoi features were collected such as number of vertices, number of edges, number of faces, face area, number of bounding edges per face, and Voronoi cell volume. Subsequently, $[\cdot]$, is itself constructed from concatenation of a series of histograms:

$$[h_{ij}] = [h_{ij}]_{k_1} \cdot [h_{ij}]_{k_2} \cdots$$

where $[\cdot]_{k_1}, [\cdot]_{k_2}$, and so on are histograms from Voronoi feature type 1, Voronoi feature type 2, and Voronoi feature type 3, and so on, respectively. Voronoi features that were included in Equation (9) are displayed in Table 1. Each Voronoi histograms were then bin-wise averaged by the total number of atoms in the original unit cell:

$$[h_{ij}] = \frac{1}{N_{\text{atoms}}} \sum_{i=1}^{N_{\text{atoms}}} [h_{ij}]_{k}^{k}$$

As there are several Voronoi feature types, we allow the machine learning models to determine naturally during the model building step which among the features are predictive for the target properties.

Given a structure coordinate data, the collection of features started from a given unit cell for input in 3D-Voronoi tessellation. To avoid the inclusion of artificial Voronoi shapes forming near the boundary wall, the unit cell was first expanded into a supercell (e.g. $3 \times 3 \times 3$) as shown in Figure 2(a). The atomic coordinates were then transformed with respect to a larger container box of arbitrary size (shown as a rectangular parallelepiped). This container box served as a global reference for the coordinates, that is, a given atom coordinate $\vec{r}$ with supercell basis vectors $\vec{a}, \vec{b}, \vec{c}$ was transformed into the container basis vectors $\vec{a}', \vec{b}', \vec{c}'$ as $\vec{r}$:

$$\vec{a}' = s_{11} \vec{a} + s_{12} \vec{b} + s_{13} \vec{c},$$

$$\vec{b}' = s_{21} \vec{a} + s_{22} \vec{b} + s_{23} \vec{c},$$

$$\vec{c}' = s_{31} \vec{a} + s_{32} \vec{b} + s_{33} \vec{c},$$

where the matrix coefficients $s_{ij}$ signifies changes in length, orientation, or both of basis vectors $\vec{a}, \vec{b}, \vec{c}$ in the new basis vectors $\vec{a}', \vec{b}', \vec{c}'$. Voronoi-feature real values were then extracted from atoms bounded only within the center subcell unit.

In general, the generated Voronoi-based histograms contain first-nearest-neighbor (1NN) information,

![Figure 1](image-url)
noting that Voronoi facets appear between 1NN atoms joined together by lines bisected by planes perpendicularly. In order to explicitly include 2NN information, partitioning was performed as well for cation and anion subset coordinates of the original unit cell. An example is shown Figure 2(b)–(e) for a test case crystal structure Pbcn LiSbWO$_6$ [47]. The overall input vector formed by combining EN and Voronoi histograms are hereafter called EN+VORO feature vector for simplicity. Voronoi tessellation was carried out using the Voro++ library [48].

**Table 1. Applied smoothing parameters for various histograms.**

| Feature              | Symbol | Histogram range | Number of bins | Smoothing factor |
|----------------------|--------|----------------|----------------|-----------------|
| Electronegativity    | EN     | 0 < EN < 5     | 50             | 2               |
| #Voronoi cell edges* | a      | 0 < v < 70     | 40             | 4               |
| Area per Voronoi cell face* | f      | 0 < e < 90     | 50             | 4               |
| #edges per Voronoi cell face* | g      | 0 < f < 50     | 30             | 4               |
| #Voronoi cell faces* | s      | 0 < a < 15     | 18             | 1               |
| Voronoi cell volume* | v      | 0 < b < 20     | 20             | 1               |
| #Voronoi cell vertices* | w      | 0 < V < 25     | 25             | 1               |
| RDF                  | RDF    | 0 < RDF < 6 Å  | 20             | 0.2 Å           |

*Features extracted around each central atom in a unit cell.

**Figure 2.** (a) Schematic workflow for the extraction procedure implemented for Voronoi tessellation of crystal structures, (b) Test case crystal structure Pbcn LiSbWO$_6$ (green spheres are Li atoms, brown spheres are Sb atoms, magenta spheres are W atoms, and red spheres are O atoms). Voronoi tessellation for LiSbWO$_6$ (c) with all atoms considered (1st nearest-neighbor (NN) information), (d) with Li atoms only (2NN information), and (e) with O atoms only (2NN information). Note: Voronoi atom centers and Voronoi edges are shown in black and gray, respectively.

### 2.3. Supervised learning details

In this work we focused on four material properties for building machine models of the form \( y = f(x) \): CE, d, BG and Ed. Given that these targets are continuous data, the supervised learning task essentially falls under a regression approach wherein an error loss function is minimized:

\[
\min_{f} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda g(f),
\]

where the first and second term are the loss and regularization function, respectively. The loss function is assumed with a squared form:

\[
L(f(x), y) = \frac{1}{2} (y - f(x))^2.
\]

Meanwhile, \( g(f) \) describes the measure of complexity or roughness of \( f \) and \( \lambda \) is a model hyperparameter that is tuned during model training and validation.

There are many regressors that can be used to evaluate the predictive performance of material descriptors and representations for numerical target material properties, each of them can lead to varying levels of out-of-sample error [49]. In this work, two distinct nonlinear regressors were chosen, the first one takes a predetermined predictor form while the second one has a form that has to be supplied by the data itself: gradient boosting regression (GBR) with decision-tree base learners and epsilon
(ε)-support vector regression (ε-SVR) with radial basis function (RBF) kernel [50, 51].

In tree-based GBR, the additive expansion of \( f(x) \) is given by the form:

\[
f(x) = \sum_{m=1}^{M} \beta_m h(x; a_m),
\]

where \( \beta_m \) represents the coefficient of base tree \( h \) with hyperparameter set \( a \). Base tree learners are fitted to residuals \( y_i - f(x_i) \), instead of \( y_i \), in a forward stage-wise manner:

\[
\Theta = \sum_{i=1}^{\ell} \frac{1}{2} (y_i - f(x_i))^2,
\]

where \( \Theta \) is the summed residuals for all \( \ell \) training data points. Taking the derivative of \( \Theta \) with respect to \( f(x_i) \) leads to the interpretation for residuals as negative gradients:

\[
y_i - f(x_i) = - \frac{\partial \Theta}{\partial f(x_i)}. \tag{18}
\]

Equation (18) enables the use of steepest-descent approach for improving prediction related to \( f(x) \). Performing a stage-wise strategy of adding the \( m \)th base tree will update \( \beta \) and \( a \):

\[
(\beta_m, a_m) = \arg\min_{\beta, a} \sum_{i=1}^{N} \Theta(y_i, f_{m-1}(x_i) + \beta h(x_i; a)) \tag{19}\]

Equation (18) then leads to the \( m \)th update of \( f(x) \):

\[
f_m(x) = f_{m-1}(x) + \beta_m h(x; a_m). \tag{20}
\]

In ε-SVR, the starting point is the familiar expression for the case of linear functions:

\[
f(x) = w \cdot x + b, \tag{21}
\]

where \( \cdot \) indicates dot product operation in \( x, w \in x \), and \( b \in \mathbb{R} \). The goal is to find \( f(x) \) that has the largest deviation \( \epsilon \) from \( y_i \), using all training data, while at the same time aiming to obtain a flat \( f(x) \) as much as possible. To extend SVR to nonlinear functions, the kernel trick is employed such as the RBF:

\[
k(x_i, x_j) = \exp\left(- \frac{||x_i - x_j||^2}{2\sigma^2}\right), \tag{22}
\]

where \( \sigma \) is a free parameter.

Data splitting for model training and validation (80% of total data-set), and testing (20% of total data-set) was instantiated randomly for 100 times at each hyperparameter setting implemented (for both GBR and SVR fitting). Model selection was based on 10-fold cross validation (CV), with the final model being selected based on the lowest test data-set error (i.e. using only excluded data points from training and validation steps). Hyperparameters for both machine models were optimized by exhaustive cross-validated search over specified range of parameters (see Table 2). All fitting tasks were performed using the scikit-learn machine learning toolkit [52]. We used the grand average RMSE as a fitting stopping criterion for hyperparameter tuning and then picked the final model instance based on:

\[
\arg\min_{t \in \{1, 2, 3, \ldots, 100\}} |e_t|, \tag{23}
\]

where \( e_t \) is the test error for fitting instance \( t \) in one hyperparameter setting.

Table 2 shows the model conditions implemented for GBR and SVR fitting. For comparison, CE and BG were also fitted using RDF feature vector consisting of histograms generated from all-atom, cation–cation, anion–anion, Li–Li, O–O, and Li–O interatomic pair contributions.

3. Results and discussion

3.1. Predictive power of formulated vector-form descriptors

Figure 3 shows the post-validation step evaluation results of the final models for DFT-CE and DFT-BG predictions with using ICSD-based test data-sets (see Table S1 in Supporting Information). Grand average RMSE and MAE\textsubscript{test} (from 100 random data-set splitting) tends to converge after 50 and 70 base tree learners, respectively, were sequentially added. Note that the present RDF feature vectors include both chemical composition (implicitly) and crystal structure information, while EN or VORO feature vectors contains as well composition (explicitly) or crystal structure information, respectively, in this study. The errors do not increase even up to 420 trees owing to the subsampling routine implemented during the training step (i.e. 20% of the training data-set were always left out during the training step). For DFT-CE, EN+VORO

### Table 2. Fitting conditions employed for GBR and SVR models.

| Model | Hyperparameter | Condition |
|-------|----------------|-----------|
| GBR   | Base tree depth | 3         |
|       | #Splitting per node | 2         |
|       | #Base trees, \( n_{\text{base}} \) \(*\) | \( \sqrt{n_{\text{base}}} \in [10, 420] \) |
|       | #Subsample features for node splitting | \( \sqrt{n_{\text{data}}} \) where \( n_{\text{data}} \) is the total number of bin features |
|       | #Subsample training data | 0.8       |
|       | Learning rate | 0.1       |
|       | Kernel coefficient | RBF       |
| SVR   | #Flatness – deviation tolerance | 1/\( h_{\text{total}} \) |
|       | Deviation cutoff, \( \epsilon \) | \( \epsilon \in [0.01, 0.60] \), \( \Delta \epsilon = 0.01 \) |

\(*\) Hyperparameters optimized by exhaustive cross-validated search over specified range of parameters.
DFT-CE, EN+VORO provides better prediction than RDF with GBR fitting by $\Delta\text{RMSE}_{\text{test}} = 21.0$ meV/atom ($\Delta\text{MAE}_{\text{test}} = 23.0$ meV/atom) and with SVR fitting by $\Delta\text{RMSE}_{\text{test}} = 19.0$ meV/atom ($\Delta\text{MAE}_{\text{test}} = 6.0$ meV/atom). For DFT-BG, EN+VORO underperforms relative to RDF: $\Delta\text{RMSE}_{\text{test}} = 0.164$ eV ($\Delta\text{MAE}_{\text{test}} = 0.154$ eV and $\Delta\text{RMSE}_{\text{test}} = 0.089$ eV ($\Delta\text{MAE}_{\text{test}} = 0.058$ eV) by GBR and SVR approach, respectively. Insets in Figure 4(a) and (c) show the converged deviance (error residuals) for training and test sets for GBR fitting, stopping at the optimal number of trees of 50 and 70 trees for DFT-CE and DFT-BG, respectively. Despite the slightly larger variance in errors for EN+VORO for DFT-BG prediction, its coefficient of determination test is still reasonable ($R^2 = 0.70$). A summary of test set prediction metrics of the final models are listed in Table 3. Improvement on generalization for EN+VORO may be realized by increasing the size of training data, given that it may have a larger learning capacity than RDF, as suggested by the larger error difference between training and test data-sets for the former (e.g. 0.673 eV vs. 0.591 eV, respectively, for $\Delta\text{RMSE}_{\text{train-test}}$ for DFT-CE by GBR with 70 trees) [53].

To further check the predictive performance of EN+VORO, fitting was also performed by GBR using

![Figure 3](image-url) Test data-set grand average errors for DFT-CE and DFT-BG by (a), (c) GBR (total: 420 trees/data split × 100 data splits = 4200 fitting instances for CE and BG, respectively) and (b), (d) SVR fitting. Test data-set error surface with respect to hyperparameter combination ($C$ and $\epsilon$) for SVR fitting for (c) DFT-CE and (d) DFT-BG; a total of 18,600 regularly-spaced hyperparameter coordinate sets (or fitting instances) each, respectively.

outperforms RDF on the average by $\Delta\text{RMSE}_{\text{ave}} = 0.053$ eV ($\Delta\text{MAE}_{\text{ave}} = 0.025$ eV). However, for DFT-BG prediction (Figure 3(b)), the reverse is true, with RDF predicting better than EN+VORO by $\Delta\text{RMSE}_{\text{ave}} = 0.110$ eV ($\Delta\text{MAE}_{\text{ave}} = 0.108$ eV). On the other hand, for DFT-CE prediction by SVR, the resulting error surface with respect to $C$ and $\epsilon$ hyperparameters is apparently stratified with peaks and trough for both EN+VORO and RDF feature vectors (Figure 3(c)). The error magnitude comparison is consistent with GBR results, that is, EN+VORO predicts more accurately than RDF by $\Delta\text{RMSE}_{\text{ave}} = 0.043$ eV/atom ($\Delta\text{MAE}_{\text{ave}} = 0.030$ eV/atom) for DFT-CE. For DFT-BG, RDF slightly predicts better than EN+VORO by $\Delta\text{RMSE}_{\text{ave}} = 0.110$ eV ($\Delta\text{MAE}_{\text{ave}} = 0.096$ eV). Error for EN+VORO is relatively more sensitive than RDF with respect to increasing $\epsilon$ (Figure 3(d) top) but is independent of $C$. RDF also has a narrower range for $\text{RMSE}_{\text{ave}}$ than EN+VORO (0.116 eV vs. 0.284 eV, respectively). The optimal SVR hyperparameter settings are \{ $C = 1900.0$, $\epsilon = 0.490$ \}_\text{DFT-CE} and \{ $C = 2600.0$, $\epsilon = 0.390$ \}_\text{DFT-BG}.

Figure 4 displays the fitting quality of the final models for DFT-CE and DFT-BG prediction with EN+VORO and RDF descriptors by GBR and SVR fitting using ICSD-based DFT data-sets with 140 compounds. For DFT-CE, EN+VORO provides better prediction than RDF with GBR fitting by $\Delta\text{RMSE}_{\text{test}} = 21.0$ meV/atom ($\Delta\text{MAE}_{\text{test}} = 23.0$ meV/atom) and with SVR fitting by $\Delta\text{RMSE}_{\text{test}} = 19.0$ meV/atom ($\Delta\text{MAE}_{\text{test}} = 6.0$ meV/atom). For DFT-BG, EN+VORO underperforms relative to RDF: $\Delta\text{RMSE}_{\text{test}} = 0.164$ eV ($\Delta\text{MAE}_{\text{test}} = 0.154$ eV and $\Delta\text{RMSE}_{\text{test}} = 0.089$ eV ($\Delta\text{MAE}_{\text{test}} = 0.058$ eV) by GBR and SVR approach, respectively. Insets in Figure 4(a) and (c) show the converged deviance (error residuals) for training and test sets for GBR fitting, stopping at the optimal number of trees of 50 and 70 trees for DFT-CE and DFT-BG, respectively. Despite the slightly larger variance in errors for EN+VORO for DFT-BG prediction, its coefficient of determination test is still reasonable ($R^2 = 0.70$). A summary of test set prediction metrics of the final models are listed in Table 3. Improvement on generalization for EN+VORO may be realized by increasing the size of training data, given that it may have a larger learning capacity than RDF, as suggested by the larger error difference between training and test data-sets for the former (e.g. 0.673 eV vs. 0.591 eV, respectively, for $\Delta\text{RMSE}_{\text{train-test}}$ for DFT-CE by GBR with 70 trees) [53].

To further check the predictive performance of EN+VORO, fitting was also performed by GBR using
and DFT-BG (0.830 eV), while it has comparable error for DFT-Ed (31.0 meV/atom) vs. EN+VORO. We note that the result on DFT-BG here is opposite to what was observed using the relatively modest data-set size from ICSD as shown in Figure 4(b); RMSE_{test} by EN+VORO is now decreased from 0.722 to 0.661 eV. This means that the machine models fitted with EN+VORO and larger data-set size show less contribution from interpreting data noise as signal during model training step (variance is now lower). Based on a computational study by Morales-Garcia et al. [54], a linear relationship between DFT-GGA and the more accurate GW method for BG of oxides can be drawn, with a systematic underestimation error of about +0.9 eV by the former. Since this +0.9 eV offset is only a systematic shift above the ideal ‘BG_{GW} = BG_{DFT-GGA}’ line, it is expected that EN+VORO would still be superior over RDF if GW method is used instead of present DFT-GGA method. However, we caution that this relationship is reasonable only for oxide-type materials. On the other hand, RMSE_{test} for DFT-BG by RDF has increased from 0.558 to 0.728 eV with the larger data-set, reflecting a lower learning capacity of RDF compared to EN+VORO. Final models for EN+VORO and RDF are summarized in Table 4, including SVR fitting results.

Above promising results validate our approach of directly binning Voronoi-based real values for descriptor vector construction for crystalline solids. It can be a larger DFT data-set of 1000 Li-containing oxide compounds randomly extracted from Materials Project [1]. Properties such as DFT-CE, density (DFT-d), DFT-BG, and decomposition energy (DFT-Ed) were chosen for supervised learning using GBR approach. Results from hyperparameter tuning with respect to the number of boosting trees are shown in the left column of Figure 5. Plots showed a typical error convergence behavior. Fitting stopping criteria of 120, 160, 300, and 430 boosting trees were chosen with respect to RMSE_{ave,test} for DFT-CE (0.305 eV/atom), DFT-d (0.701 g/cm³), DFT-BG (0.792 eV), and DFT-Ed (31 meV/atom), respectively (final models shown in the right column of Figure 5). RMSE_{ave,test} for RDF is noted to be higher for DFT-CE (0.451 eV/atom), DFT-d (0.794 g/cm³) and DFT-BG (0.830 eV), while it has comparable error for DFT-Ed (31.0 meV/atom) vs. EN+VORO. We note that the result on DFT-BG here is opposite to what was observed using the relatively modest data-set size from ICSD as shown in Figure 4(b); RMSE_{test} by EN+VORO is now decreased from 0.722 to 0.661 eV. This means that the machine models fitted with EN+VORO and larger data-set size show less contribution from interpreting data noise as signal during model training step (variance is now lower). Based on a computational study by Morales-Garcia et al. [54], a linear relationship between DFT-GGA and the more accurate GW method for BG of oxides can be drawn, with a systematic underestimation error of about +0.9 eV by the former. Since this +0.9 eV offset is only a systematic shift above the ideal ‘BG_{GW} = BG_{DFT-GGA}’ line, it is expected that EN+VORO would still be superior over RDF if GW method is used instead of present DFT-GGA method. However, we caution that this relationship is reasonable only for oxide-type materials. On the other hand, RMSE_{test} for DFT-BG by RDF has increased from 0.558 to 0.728 eV with the larger data-set, reflecting a lower learning capacity of RDF compared to EN+VORO. Final models for EN+VORO and RDF are summarized in Table 4, including SVR fitting results.

Above promising results validate our approach of directly binning Voronoi-based real values for descriptor vector construction for crystalline solids. It can be

### Table 3. Comparison of predictive accuracy between EN+VORO and RDF-based final machine models with ICSD-based DFT calculated data-set (140 compounds).

| Target property | Feature vector | Model | Test set RMSE | Test set MAE |
|-----------------|----------------|-------|---------------|--------------|
| DFT-CE (eV/atom) | EN+VORO | GBR | 0.316 | 0.258 |
|                 |         | SVR | 0.247 | 0.196 |
|                 | RDF    | GBR | 0.337 | 0.281 |
|                 |         | SVR | 0.266 | 0.202 |
| DFT-BG (eV)     | EN+VORO | GBR | 0.722 | 0.587 |
|                 |         | SVR | 0.497 | 0.387 |
|                 | RDF    | GBR | 0.558 | 0.433 |
|                 |         | SVR | 0.408 | 0.329 |

Figure 4. Fitting quality of final models: (a) CE with GBR, (b) CE with SVR, (c) BG with GBR, and (d) BG with SVR. Insets in (a) and (c) show deviance (error residuals) plots terminating at the optimal number of base tree learners for GBR fitting (50 trees for CE and 70 trees for BG, respectively). Optimal hyperparameters (C and ε) for SVR fitting are indicated in (b) and (d).
coordination and environments (see Refs. [32,33]). In the case of RDF, atomic/elemental features are implicit only to the atom type pairings used for generating distance features and the 3D spatial information in terms of the relative positions of more than two atoms is mostly lost as well.

3.2. Feature importance analysis from GBR modeling

Figure 6 shows the variable importance (VI) score plots (color gradient scaled to 100) of various feature bins constituting EN+VORO descriptors. The horizontal axis denotes the histogram bin index number arranged in increasing order from left to right with respect to the subset feature real value. The vertical axis indicates the description for each subset features.

argued that the reason why EN+VORO generally out-performs RDF is that the former contains richer information (data signal over noise ratio is larger) and a larger learning capacity owing to the explicit inclusion of atomic/elemental features and the Voronoi features that are primarily generated by high-symmetry directions between lattice points of atom centers. That is why Voronoi-based information can be effectively used as criteria to reasonably define cutoffs for local atom coordination and environments (see Refs. [32,33]). In the case of RDF, atomic/elemental features are implicit only to the atom type pairings used for generating distance features and the 3D spatial information in terms of the relative positions of more than two atoms is mostly lost as well.

Table 4. Comparison of predictive accuracy between EN+VORO- and RDF-based final machine models using Materials Project data-set (1000 compounds).

| Target property | Feature vector | Model  | RMSEtest | MAEtest |
|-----------------|----------------|--------|----------|---------|
| DFT-CE (eV/atom) | EN+VORO        | GBR    | 0.210    | 0.155   |
|                 |                | SVR    | 0.274    | 0.196   |
|                 | RDF            | GBR    | 0.381    | 0.284   |
|                 |                | SVR    | 0.410    | 0.318   |
| DFT-d (g/cm³)   | EN+VORO        | GBR    | 0.575    | 0.436   |
|                 |                | SVR    | 0.532    | 0.404   |
|                 | RDF            | GBR    | 0.653    | 0.479   |
|                 |                | SVR    | 0.603    | 0.444   |
| DFT-BG (eV)     | EN+VORO        | GBR    | 0.661    | 0.515   |
|                 |                | SVR    | 0.668    | 0.541   |
|                 | RDF            | GBR    | 0.728    | 0.579   |
|                 |                | SVR    | 0.730    | 0.590   |
| DFT-Ed (meV/atom) | EN+VORO   | GBR    | 26.0     | 22.0    |
|                 |                | SVR    | 31.0     | 26.0    |
|                 | RDF            | GBR    | 27.0     | 21.0    |
|                 |                | SVR    | 30.0     | 25.0    |
up to ~25). All 2NN features for Li-cations and most of O-anions contribute as well (VI score up to ~40).

Based from above observations on feature importance, 2NN Voronoi features are consistently determined as good predictors in comparison to their 1NN counterparts. EN features are also confirmed to be good predictors. The low VI scores of 1NN Voronoi features may be explained by fluctuation effect from artificial increase of histogram bin counts because of Voronoi cell faces contributing significantly (VI score reaching as much as ~40). For the O-anions features, face area and cell volume are descriptive (VI score = ~30).

For DFT-BG predictions, several bins from the EN histogram also show high VI scores (up to 50 and higher). Again, no notable 1NN bin features appear to be meaningful. Similar with DFT-CE, 2NN bins from face area and volume from all-cations and all-anions are contributing towards DFT-BG prediction (VI score up to ~25). All 2NN features for Li-cations and most of O-anions contribute as well (VI score up to ~40).

Based from above observations on feature importance, 2NN Voronoi features are consistently determined as good predictors in comparison to their 1NN counterparts. EN features are also confirmed to be good predictors. The low VI scores of 1NN Voronoi features may be explained by fluctuation effect from artificial increase of histogram bin counts because of Voronoi cell faces contributing significantly.
that are actually not physically meaningful (i.e. faces that are not formed by nearest-neighbor information or formation of chemical bonding), causing ambiguity when differentiating crystal structures [55]. This is supported by the analysis of Figure 7(a) which displays the linear relationship between number of vertices vs. counted small face areas (<0.2 average face area unit) for DFT-optimized structures related to Figures 3 and 4. It is clear that 1NN features tend to have smaller areas with large number of vertices than 2NN features. Intuitively, this causes the resulting bin counts to be spread out leading to a histogram feature vector with the primary peak broadened and shifted towards the low value region (see Figure 7(b)). In the case of 2NN features, the histogram shape shows a relatively sharp primary peak.

4. Conclusions

A generalized histogram-based representation for crystalline solids, which is invariant to atom ordering, cell periodicity, and number of atoms, was constructed by binning atomic/chemical property data and Voronoi-tessellation features. The formulated material descriptor vector showed a better predictive power and generalization performance relative to RDF-based descriptor vector for DFT-calculated properties such as cohesive energy, density, band gap, and decomposition energy. From variable importance analysis, it was confirmed that binned atomic/chemical property and 2NN Voronoi cell geometric property data are good predictors.

Acknowledgments

R.J. thanks the JST Precursory Research for Embryonic Science and Technology (PRESTO) program for the financial support. DFT Computations were mainly carried out at the Information Technology Center of Nagoya University (CX400). Visualization for crystal structures was made using the VESTA software [56]. Voronoi cell figures were created using POV-Ray software [57], data plot figures were generated using python Matplotlib [58].

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by ‘Materials research by Information Integration’ Initiative (MiT†); the Development Program of the Japan Science and Technology Agency (JST) under ‘Advanced Materials Informatics through Comprehensive Integration among Theoretical, Experimental, Computational, and Data-centric Sciences’ research area and by ‘Elements Strategy Initiative to Form Core Research Center’ (Since 2012); Ministry of Education Culture, Sports, Science and Technology (MEXT); JST Precursory Research for Embryonic Science and Technology (PRESTO) program for the financial support.

References

[1] Jain A, Hautier G, Moore CJ, et al. A high-throughput infrastructure for density functional theory calculations. Comp Mater Sci. 2011;50:2295–2310.
[2] Saal JE, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). Jom–Us. 2013;65:1501–1509.
[3] Taylor RH, Rose F, Toher C, et al. A RESTful API for exchanging materials data in the AFLLOWLIB.org consortium. Comp Mater Sci. 2014;93:178–192.
[4] Cheng L, Assary RS, Xu QH, et al. Accelerating electrolyte discovery for energy storage with high-throughput screening. J Phys Chem Lett. 2015;6:283–291.
[5] Fujimura K, Seko A, Koyama Y, et al. Superionic conductors based on first-principles calculations and machine learning algorithms. Adv Energy Mater. 2013;3:980–985.
[6] Ghandbeigi L, Harada JK, Lettiere BR, et al. Performance and resource considerations of Li-ion battery electrode materials. Energy Environ Sci. 2015;8:1640–1650.
[7] Hautier G. Prediction of new battery materials based on ab initiocomputations. Aip Conf Proc. 2016;1765:020009(1–10).
[8] Jalem R, Kimura M, Nakayama M, et al. Informatics-aided density functional theory study on the Li ion transport of tavorite-type LiMTO₅F (M¹⁺–T⁵⁺, Ni²⁺–T⁶⁺). J Chem Inf Model. 2015;55:1158–1168.
[9] Nishijima M, Ootani T, Kamimura Y, et al. Accelerated discovery of cathode materials with prolonged cycle life for lithium-ion battery. Nat Commun. 2014;5:4553(1–7).
[10] Gautier R, Zhang XW, Hu LH, et al. Prediction and accelerated laboratory discovery of previously unknown 18-electron ABX compounds. Nat Chem. 2015;7:308–316.
[11] Landrum GA, Penzotti JE, Putta S. Machine-learning models for combinatorial catalyst discovery. Meas Sci Technol. 2005;16:270–277.
[12] Ma XF, Li Z, Achenie LEK, et al. Machine-learning-augmented chemisorption model for CO₂ reduction kinetic screening. J Phys Chem Lett. 2015;6:3528–3533.
[13] García-Cañadas J, Adkins NJE, McCain S, et al. Accelerated discovery of thermolectric materials: combinatorial facility and high-throughput measurement of thermoelectric power factor. ACS Comb Sci. 2016;18:314–319.
[14] Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. Phys Rev Lett. 2015;115:205901(1–5).
[15] Dey P, Bible J, Datta S, et al. Informatics-aided bandgap engineering for solar materials. Comp Mater Sci. 2014;83:185–195.
[16] Isayev O, Fourches D, Muratov EN, et al. Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem Mater. 2015;27:735–743.
[17] Sanvito S, Oses C, Xu J, et al. Accelerated discovery of new magnets in the Heusler alloy family. Sci Adv. 2017;3:e1602241(1–9).
[18] Ghiringhelli LM, Vybihal J, Levchenko SV, et al. Big Data of Materials Science: Critical Role of the Descriptor. Phys Rev Lett. 2015;114:105503(1–5).
Jalem R, Aoyama T, Nakayama M, et al. Multivariate ICSD, Inorganic Crystal Structure Database. 2006.

Csonka GI, Perdew JP, Ruzsinszky A, et al. Assessing Kresse G, Joubert D. From ultrasoft pseudopotentials Kresse G, Furthmuller J. Efficient iterative schemes for Isayev O, Oses C, Toher C, et al. Universal fragment Yang LS, Dacek S, Ceder G. Proposed definition of Voronoi G. Recherches sur les paralléloèdres primitifs Seko A, Hayashi H, Nakayama K, et al. Representation Schutt KT, Glawe H, Brockherde F, et al. How to Rupp M, Tkatchenko A, Muller KR, et al. Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett. 2012;108:058301(1–5).

Schütz KT, Glave H, Brockherde F, et al. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. Phys Rev B. 2014;89:205118(1–5).

Seko A, Hayashi H, Nakayama K, et al. Representation of compounds for machine-learning prediction of physical properties. Phys Rev B. 2017;95:144110(1–11).

Bartók AP, De S, Poelking C, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3:e1701816(1–8).

Bartók AP, Payne MC, Kondor R, et al. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett. 2010;104:136403(1–5).

Voronoi G. Recherches sur les parallélôédres primitifs [Research on primitive parallelhedrons]. J Reine Angew Math. 1908;134:198–287.

Voronoi G. Recherches sur les parallélôédres primitifs [Research on primitive parallelhedrons]. J Reine Angew Math. 1909;136:67–182.

Yang LS, Dacek S, Ceder G. Proposed definition of crystal substructure and substructural similarity. Phys Rev B. 2014;90:054102(1–9).

Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun. 2017;8:15679(1–12).

Kresse G, Furthmuller J. Efficient iterative schemes for \textit{ab initio} total-energy calculations using a plane-wave basis set. Phys Rev B. 1996;54:11169–11186.

Kresse G, Joubert D. From ultrasoft pseudopotentials to the projector augmented-wave method. Phys Rev B. 1999;59:1758–1775.

Csonka GI, Perdew JP, Ruzsinszky A, et al. Assessing the performance of recent density functionals for bulk solids. Phys Rev B. 2009;79:155107(1–32).

ICSD, Inorganic Crystal Structure Database. 2006. Available from: http://icsd.fizkarlsruhe.de/icsd/.

Payne MC, Teter MP, Allan DC, et al. Iterative minimization techniques for \textit{ab initio} total-energy calculations: molecular dynamics and conjugate gradients. Rev Mod Phys. 1992;64:1045–1097.

Liechtenstein AI, Anisimov VI, Zaanen J. Density-functional theory and strong interactions: orbital ordering in Mott-Hubbard insulators. Phys Rev B. 1995;52:R5467–R5470.

Dudarev SL, Botton GA, Savrasov SY, et al. Electron-energy-loss spectra and the structural stability of nickel oxide: an LSDA+U study. Phys Rev B. 1998;57:1505–1509.

Hedin L. New method for calculating the one-particle green’s function with application to the electron-gas problem. Phys Rev. 1965;139:A796–A823.

Aulbur WG, Jonsson L, Wilkins JW. Quasiparticle calculations in solids. Solid State Phys. 2000;54:1–218.

Pauling L. The nature of the chemical bond. IV. The energy of single bonds and the relative electronegativity of atoms. J Am Chem Soc. 1932;54:3570–3582.

Shannon RD. Revisited effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. Acta Cryst. 1976;A32:751–767.

KanL L. On dimensionality and sample size in statistical pattern classification. Pattern Recognit. 1971;3:225–234.

O’Keeffe M. A proposed rigorous definition of coordination number. Acta Cryst. 1979;A35:772–775.

Louer M, Le Roux C, Dubac J. Ab initio structure determination from X-ray powder diffraction data of tetraaquabis(mth III) trflate obtained from the nonahydrate. Chem Mater. 1997;9:3012–3016.

Rycroft CH. VORO plus plus: a three-dimensional Voronoi cell library in C plus plus. Chaos. 2009;19:041111(1–1).

Fabre FA, Hutchison L, Huang B, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. J Chem Theory Comput. 2013;13:5255–5264.

Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–1232.

Smola AJ, Scholkopf B. A tutorial on support vector regression. Stat Comput. 2004;14:199–222.

Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–2830.

Vapnik V, Levin E, Le Cun Y. Measuring the VC-dimension of a learning machine. Neural Comput. 1994;6:851–876.

Morales-García Á, Valero R, Illas F. An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations. J Phys Chem C. 2017;121:18862–18866.

Brostow W, Chybicki M, Laskowski R, et al. Voronoi polyhedra and Delaunay simplices in the structural analysis of molecular-dynamics-simulated materials. Phys Rev B. 1998;57:13448–13458.

Momma K, Izumi F. VESTA: a three-dimensional visualization system for electronic and structural analysis. J Appl Cryst. 2008;41:653–658.

Persistence of Vision Raytracer. Version 3.6. [Computer software]. Persistence of Vision Pty. 2004. Available from: http://www.povray.org/download/.

Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9:90–95.