npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Topological representations of crystalline compounds for the machine-learning prediction of materials properties

Yi Jiang[1], Dong Chen[1,2], Xin Chen[1], Tangyi Li[1], Guo-Wei Wei[2]✉ and Feng Pan[1]✉

Accurate theoretical predictions of desired properties of materials play an important role in materials research and development. Machine learning (ML) can accelerate the materials design by building a model from input data. For complex datasets, such as those of crystalline compounds, a vital issue is how to construct low-dimensional representations for input crystal structures with chemical insights. In this work, we introduce an algebraic topology-based method, called atom-specific persistent homology (ASPH), as a unique representation of crystal structures. The ASPH can capture both pairwise and many-body interactions and reveal the topology-property relationship of a group of atoms at various scales. Combined with composition-based attributes, ASPH-based ML model provides a highly accurate prediction of the formation energy calculated by density functional theory (DFT). After training with more than 30,000 different structure types and compositions, our model achieves a mean absolute error of 61 meV/atom in cross-validation, which outperforms previous work such as Voronoi tessellations and Coulomb matrix method using the same ML algorithm and datasets. Our results indicate that the proposed topology-based method provides a powerful computational tool for predicting materials properties compared to previous works.

npj Computational Materials (2021)7:28 ; https://doi.org/10.1038/s41524-021-00493-w

INTRODUCTION

Advances in materials science are typically slow and arduous[1], and thus is particularly challenging to meet the increased demand for material characterization[2]. Because the number of possible materials is estimated to be as high as a googol (10^{100})[3], there is an urgent need for innovative methods and techniques in materials research. To accelerate the development of new materials, high-throughput computing methods have been proposed in recent years, especially the density functional theory (DFT) which can predict the properties of both experimental and hypothetical inorganic compounds[4,5]. The combination of both experiments and computer simulations has proven to be a powerful approach to reduce the time and cost of materials design and has been widely used in Li-ion batteries, electrocatalysis, thermoelectrics, and structural alloys. It has also promoted the establishment of large and high-quality open databases such as Materials Project (MP)[6], Open Quantum Materials Database (OQMD)[7], the Automatic Flow of Materials Discovery Library[8], and MaterialGo[9]. While DFT approaches are powerful, they are inefficient and prohibitively expensive for heavier elements, strongly correlated electrons, and large molecules. In addition, all physical methods, including DFT, are not designed to deal with massive and diverse datasets in materials sciences.

With the development of large databases and improved algorithms, machine learning (ML) has emerged as a great promising approach in the research of inorganic crystal structures and molecules. It offers a revolutionary tool for rapidly estimating the results of DFT calculations or experimental data by creating prediction models from databases. ML algorithms aim to optimize the generalization performance of models. ML might be generally split into three main categories: supervised learning, unsupervised learning, and reinforcement learning. During the process of supervised learning, systems are exposed to large amounts of labeled data to find the unknown function that can extrapolate the result of unlabeled data. In contrast, unsupervised learning tasks with identifying patterns in data and trying to looks for unlabeled data that can be grouped by similarities. Reinforcement learning aims to learn good policies for sequential decision problems by optimizing a cumulative future reward signal[10]. ML has been successfully applied to predict materials properties including formation energy[11–13], bandgap[14–19], thermal conductivity[20–22], and elastic modulus[23–25]. It can be used to create atomic potential[26,27], screen functional materials[28–30], and analyze complex reaction networks[31].

Descriptors or features, as a pivotal ingredient of a ML model, provide a representation of each molecule in a data set[32]. A poor representation that is either unable to reduce the complexity of the data or unable to maintain vital material information will inevitably lead to large prediction errors. More specifically, descriptors in materials science should capture the information that could distinguish between different atomic and crystal environments[33]. Several different strategies have already been proposed to extract the quantitative representations of crystal materials. Potential energy is predicted by the transformation of pairwise distance models[34], which only work for a fixed number of atoms and are not unique under the permutation of atoms. Several models rely on the dataset of compounds with the same stoichiometry or the same structure[11,35,36]. However, to fully cover highly diverse compositions and structures in crystal materials, a rotational, transnational, and scale-invariant representation is needed to empower ML models. Faber et al.[37] proposed three generalized Coulomb matrix (CM) approaches. One approach considers full Coulomb interactions between two atoms in a lattice setting. The second one models atomic electrostatic interactions in the unit cell and its nearest neighbor environment. In the last approach, they replace the Coulomb interaction by a periodicity potential with respect to the lattice vectors. Their approaches achieved 0.37 eV/

[1]School of Advanced Materials, Peking University, Shenzhen Graduate School, Shenzhen, PR China. [2]Department of Mathematics, Michigan State University, East Lansing, MI, USA. ✉email: weig@msu.edu; panfeng@pkusz.edu.cn

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

Table 1. The overall performance of formation enthalpy predictor with different feature extraction method on cross-validation (CV) tasks.

Method[a]	R^2	RMSE (eV/atom)	MAE (eV/atom)
ASPH + Comp[b]	0.986 (0.000)	0.119 (0.000)	0.061 (0.000)
VT	0.983 (0.000)	0.129 (0.000)	0.067 (0.000)
ASPH	0.970 (0.000)	0.174 (0.000)	0.103 (0.000)
Betti-0ASPH	0.969 (0.000)	0.177 (0.000)	0.108 (0.000)
CM-sine	0.930 (0.000)	0.169 (0.000)	0.267 (0.000)

[a] The average performance is reported and standard deviations are given in square brackets.

[b] Comp refers to composition-based attributes.

atom mean absolute error (MAE) in predicting the formation energies of new structures. Schütt et al.[38] constructed a ML model to predict the density of states at the Fermi energy based on the crystal representation called partial radial distribution function which is invariant under translation, rotation, and the choice of the unit cell. Ward et al.[39] applied the standard random forest (RF) to predict the formation energy based on features derived from Voronoi tessellations to represent structural properties and atomic properties. This model achieves an MAE of 80 meV/atom in cross-validation for a dataset of 435,000 formation energies. These feature engineering-based ML models with structural descriptors have achieved exceptional accuracy prediction. Xie et al.[40] proposed another idea to construct features by crystal graph convolutional neural network, which is invariant for unit cell choices and achieves a high prediction accuracy of DFT calculations on many properties. However, neural networks are well-known "black box" and involve too many parameters. After intensive training, the final predictors are hard to interpret physically. In contrast, topology considers the global connectivity of various components in a space and studies isolated entities, rings, and higher dimensional faces within the space[41]. It turns out that traditional topology gives rise too much geometric reduction to provide useful description of crystal structures[42]. Persistent homology, however, is able to bridge geometrical shape analysis and topological characterization by embedding multiscale geometric information into topological invariants[43]. It generates a nested family of topological spaces by varying a filtration parameter, which results in topological invariants of various dimensions, namely, isolated components, circles, and cavities, corresponding to Betti-0, Betti-1, and Betti-2, respectively. Molecule-level persistent homology neglects chemical and biological properties, element-specific persistent homology has been proposed to retain crucial biological information[44–46]. This method has been applied to represent organic molecular and biomolecular properties[44–46]. The successful application in biomolecules motivates us to utilize persistent homology to represent crystal compounds for predicting their physical properties.

In this work, we propose a enhanced approach for predicting properties of crystalline materials using a topological representation derived from persistent homology. The direct application of persistent homology without considering atomic diversity and crystal periodicity is not suitable for crystal property predictions. Therefore, we introduce atom-specific persistent homology (ASPH) to extract atom-specific crystal information for representing crystal structures in ML. ASPH offers a variety of atom-specific topological fingerprints in the crystal cell and adapts this representation to periodic systems. In this work, we employ a large dataset from Inorganic Crystal Structure Database (ICSD)[47] and OQMD to compare our proposed method against existing methods in the literature (i.e., Voronoi tessellations and CM) via cross-validation. Combined with composition-based attributes, our method achieved excellent results with the mean absolute error as low as 61 meV/atom. Moreover, to understand the limitations of our topology-based ML method, we analyze the outliers with large errors in our predictions with respect to DFT calculations. We explore the types of compounds that are more likely to cause large deviations in our predictions.

RESULTS AND DISCUSSION

General performance

In Table 1, we compare the performance of the proposed topology-based method with those of Voronoi tessellations[39] and Coulomb Matrix-based method in the literature[37] using the same machine learning algorithm and the same set of hyperparameters. The scatter plots of the predictions and performance for different prediction methods are shown in Fig. 1. ASPH refers to the method using topological invariants Betti-0, Betti-1, and Betti-2. Similarly, Betti-0 ASPH only use Betti-0 for feature generations. The coefficients of determination (R^2), the root mean squared error (RMSE), and the mean absolute error (MAE) for tenfold cross-validations repeated 20 times are given for various methods. An MAE of 61 meV/atom is achieved by ASPH combined with composition-based features, whose performance is better than those of Voronoi tessellations and CM modified by sine matrix approximation. Our mean prediction error is also lower than the error of DFT approximation to experimentally measured formation enthalpies. Moreover, we find that the MAE from Betti-0 ASPH is slightly larger than the one from ASPH. We found that it is necessary to add topological invariants Betti-1 and Betti-2 to feature because they can capture the many-body interactions and reflect the symmetry of crystal structure. To better illustrate our findings, we consider a highly symmetrical structure NaCl in Fig. 2. With the radius of filtration increasing, when there is only one component, which means all atoms in neighbor are in connection, all 1-simplex in the point cloud directly turn into 3-simplex. Therefore, there is no Betti-1 in its point cloud. In addition to that, we find that if using only Betti-0 representations, we can also achieve good performance. Apart from having a lower MAE, the model created by our method has a better result on R^2 and RMSE as well. Overall, for our method, predicted values of 28195 (88.3%) materials are within 25% of the DFT-calculated values, and only 53 (0.17%) predicted values have errors over 1 eV/atom. In contrast, for Voronoi tessellations and CM-sine, 27,948 and 22,404 predicted values are within 25% accuracy of computed values while 66 and 315 predicted values have errors over 1 eV/atom, respectively. Supplementary Fig. S1 shows the learning curves of multiple model prediction performance with respect to the training size. It indicates that topological attributes provide important information about the crystal materials and improve accuracy compared with composition-only features. Moreover, our model prediction accuracy is higher than other methods as the amount of the training data increases. Because the topological properties provide unique information when there are multiple structures with the same composition.

Systematic errors analysis

To evaluate the scalability of our topological based model, we analyze the absolute prediction error in the tenfold cross-validations of our method for each compound in the dataset. We select the 638 compounds with the highest 2% prediction error values (i.e., above 0.336 eV/atom) to understand the set of compounds that are difficult to be accurately predicted. From Fig. 3a, it is clear that many of these compounds have positive formation enthalpies, suggesting they are thermally unstable. Therefore, the unstable compounds are most likely outliers and their experimental values are subject to large errors. It is likely true that the original DFT calculations were also unreliable for these compound.

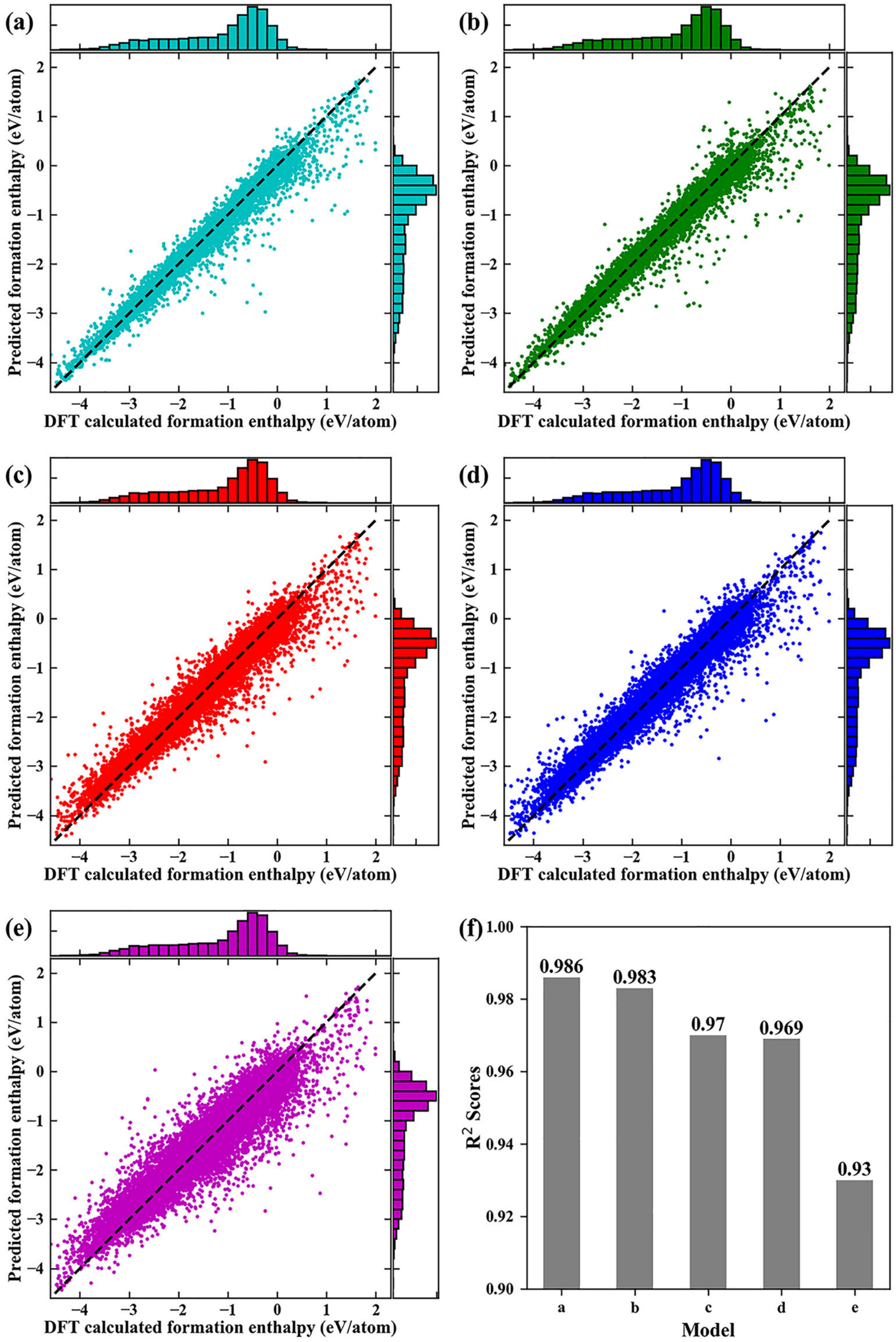

Fig. 1 Comparison of DFT-calculated formation enthalpy and predicted formation enthalpy (eV/atom) for different methods. a ASPH + Comp, **b** Voronoi tessellations, **c** ASPH, **d** Betti-0 ASPH, and **e** CM-sine. The top and right subfigure is the distribution of calculated data from ICSD and predicted data, respectively. **f** R^2 results for different models. All prediction data are the average performance obtained from tenfold cross-validations with 20 repetitions.

We also found that compounds that contain elements that occur less frequently in our dataset are more likely to appear in the set with higher prediction errors. Figure 3b shows the comparison of the probability of finding a given compound with a specific element in dataset [*P*(Dataset)] and the ratio between the probability of finding that element in the set of 2% highest prediction errors [*P*(Worst)] and the probability of finding the same element in our entire dataset. The least frequently occurring elements in our dataset such as Pu, Re, Ta, and Eu, have a higher probability to appear in the set with top 2% highest prediction errors. From the above results, it is clear that our model is less predictive for molecules having rarely occurring elements in the database. This is true for ML-based methods in general.

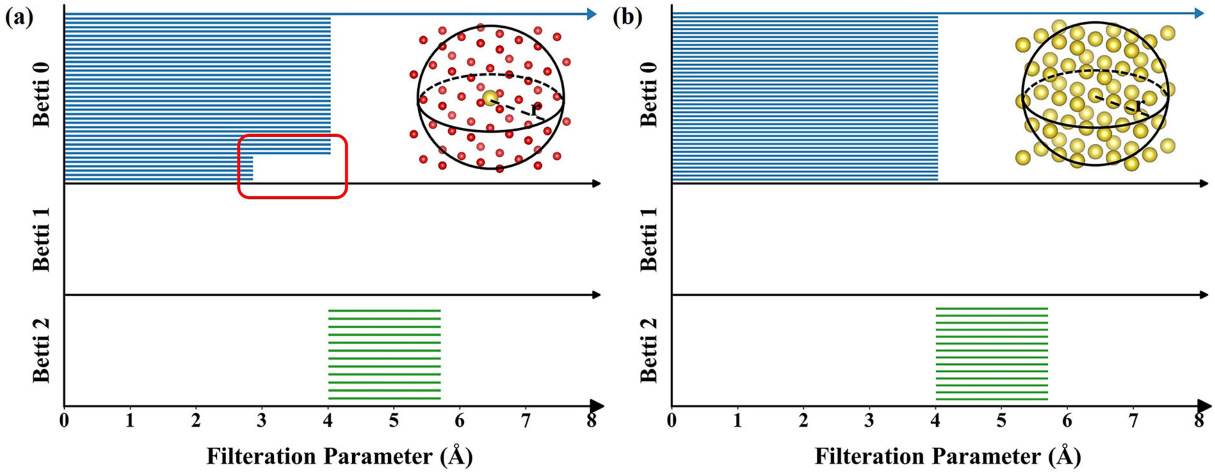

Fig. 2 Illustration of persistence barcodes of NaCl crystal specific central-atom persistent homology. The central atom is Na with the surrounding atom is Cl (**a**) and is Na (**b**), respectively. Charts from top to bottom are Betti-0, Betti-1, and Betti-2 barcodes, respectively. The point cloud of each barcode is displayed.

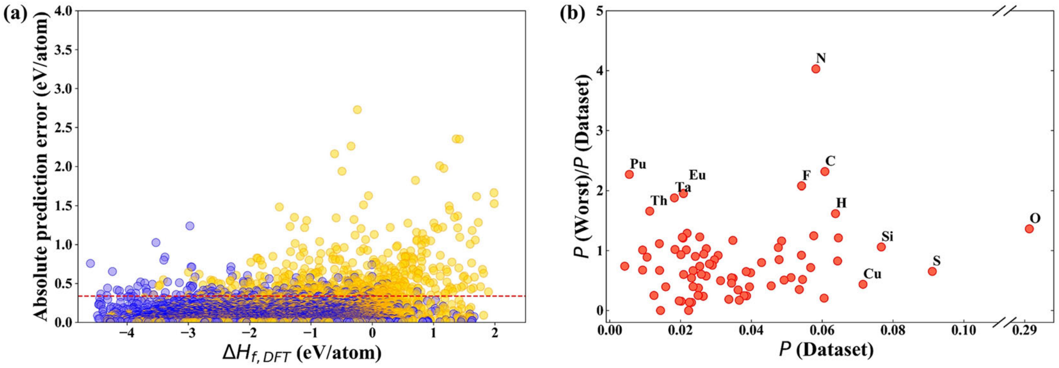

Fig. 3 Analysis of ML model prediction error. a An illustration of the DFT-calculated formation enthalpies of various compounds and their absolute prediction errors in 10-fold cross-validations. The red dashed line refers to the error value to which 98% of absolute prediction errors are smaller or equal. The yellow dots indicate that the ML predicted values are less than those of the DFT while the blue dots are the opposite. **b** An illustration of the occurrence probabilities of various elements in the whole dataset *P*(Dataset) vs. the ratios of their occurrence probabilities in the set of 2% highest prediction errors [*P*(Worst)] over *P*(Dataset).

There are three elements, N, C, and F, that occur frequently in the entire dataset but are some of the worst predicted compounds. In 638 compounds with high prediction errors, there are 217 compounds that contain N, C or F. We found that these elements occur in association with other rarely-occurring elements, such as IrN, TlF_3, IrC_4, which makes their accurate predictions very difficult. Therefore, the [*P*(Worst)] of these three elements will also be relatively high. Additionally, a few compounds with element F have positive formation enthalpies (3/1730) while other compounds contain element F have negative formation enthalpies. Therefore, the predicted formation enthalpies of these three outlier compounds with element F are negative and their absolute prediction errors are larger than 1 eV/atom.

$AlPO_4$ (ICSD #162670) structure shown in Supplementary Fig. S2 exhibits the worst prediction accuracy with an absolute prediction error (APE) of 2.73 eV/atom. Compared with other stereoisomerisms in our dataset, we find that most of DFT-calculated formation enthalpy values are <-2.90 eV/atom while there are two outliers with $\Delta H_f = -0.24$ eV/atom, -0.61 eV/atom and APE $=$ 2.73 eV/atom, 2.16 eV/atom, respectively. An illustration of DFT-calculated enthalpy values and the APE of $AlPO_4$ is shown in Supplementary Fig. S3. For similar reasons, $BaTiO_3$ (ICSD #109327), BN (ICSD #27986), CaO (ICSD #261847), VS_2 (ICSD #68713), etc. also have high errors. From our analysis of their failures, we conclude that our model does not give accurate predictions of formation enthalpy values for stereoisomerisms having a diverse formation enthalpy distribution. A possible reason is that the procedure of processing topological information is oversimplified so that ML algorithm does not do a good job in differentiating stereoisomerisms.

Contrary to geometry that widely used in crystal structure descriptors, topology is rarely implemented in quantitative analysis of materials science. In this work, we propose atom-specific persistent homology (ASPH) and apply it to material science analysis via machine learning (ML) models. Unlike high-level abstraction of conventional topology, the proposed ASPH embeds multiscale geometric information into topological invariants with chemical insights. It can effectively extract unique features such as independent components, loops, and cavities. More specifically, independent components are associated bond lengths of pairwise interactions, while loops and cavities capture many-body interactions.

Furthermore, our ASPH can deal with crystalline compounds which have structural periodicity and elemental diversity. Extensive experiment shows that our model provides a reliable estimation of DFT calculations using around 30,000 training data with diverse structural types and compositions. Moreover, it offers a more accurate prediction in cross-validations than previous

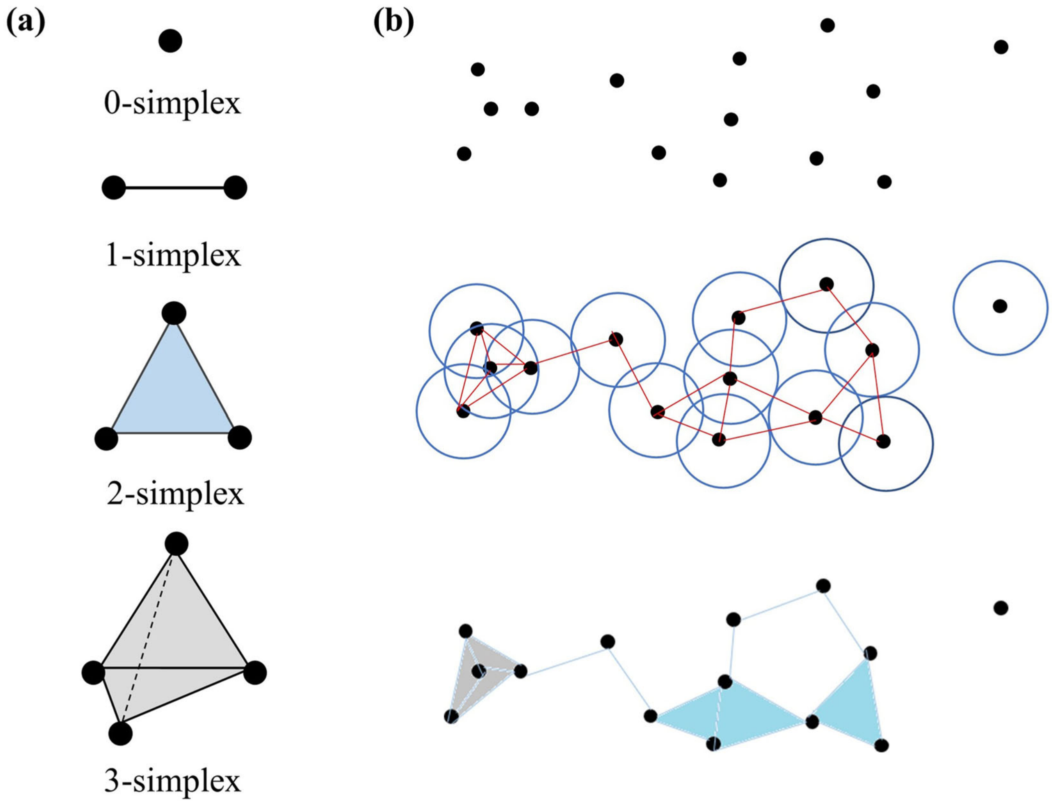

Fig. 4 Basic simplexes and simplicial complex construction in a given radius of filtration. a From top to bottom an example of a 0-simplex, 1-simplex, 2-simplex, and 3-simplex. **b** The construction of simplicial complex. There are one 0-simplex, six 1-simplexes, two 2-simplexes and one 3-simplex.

methods do[37,39]. Its applicability extends to all space groups and a great majority of elements. Finally, the success of this method enables us to discover new materials with desirable properties significantly faster and cheaper.

METHODS

Simplex and simplicial complex

Topological data analysis uses a simplices and simplicial complexes for the description of complex shapes, which are mathematically and computationally easier to process than their original counterparts. A set of $k+1$ affinely independent points in $\mathbb{R}^k$ is a k-simplex denoted by σ^k which can be represented by $\{v_0, v_1, \cdots, v_k\}$ and each v_i is called a vertex of the simplex. Specifically, a 0-simplex is a vertex, a 1-simplex is a line segment, a 2-simplex is a triangle and a 3-simplex is a tetrahedron, as shown in Fig. 4. A subset of the $k+1$ vertices of a k-simplex with $m+1$ vertices forms a convex hull in a lower dimension and is called an m-face of the k-simplex. An m-face is proper if $m < k$. A simplicial complex is a set of simplices which are convex hulls of affinely independent points. More specifically, a simplicial complex is a finite collection of simplices $X = \{\sigma_i\}_i$ satisfying that the intersection of any two simplices in X is either empty set or a common face of the two and all the faces of a simplex in X is also in X. The collection of all k-simplices in X is denoted X_k. The dimension of a simplicial complex is the highest dimension of its simplices.

Homology

For a simplicial complex K, a k-chain c_k of K is the sum of the form of k-simplices in K, and k is not greater than dimension of K, and is defined as $c_k = \Sigma a_i \sigma_i$ where σ_i is the k-simplices and a_i is coefficients. Generally, a_i can be set as elements of a field such as $\mathbb{R}$, $\mathbb{Q}$, or $\mathbb{Z}_n$. For simplicity, it is commonly chosen to be $\mathbb{Z}_2$. The group of k-chains in K, denoted C_k, with operation of modulo 2 addition can form an Abelian group $(C_k, \mathbb{Z}_2)$. So we can extend the definition of the boundary operator to chains, showed in Eq. (1).

The boundary operator applied to a k-chain c_k is defined as

$$\partial_k \sigma_k = \sum a_i \partial_k \sigma_i, \tag{1}$$

where σ_i's are k-simplices. The boundary operator is a map from C_k to C_{k-1}, which is also named boundary map for chains. Note that operator ∂_k satisfies the property that $\partial_k \partial_{k+1} = \varnothing$ for any $(k+1)$-simplex σ following the fact that any $(k-1)$-face of σ is contained in exactly two k-faces of σ. The chain complex is defined as a sequence of chains connected by boundary maps with decreasing dimensions and is represented as

$$\cdots \to C_n(K) \xrightarrow{\partial_n} C_{n-1}(K) \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0. \tag{2}$$

In other words, through the application of two boundary operations, the k-chain is mapped to an empty set $\partial_k \partial_{k+1} = \varnothing$, we can define k-cycle group and k-boundary group which are the subgroups of C_k as kernel and image of ∂_k and ∂_{k+1}, respectively,

$$Z_k = \mathrm{Ker}\, \partial_k = \{c \in C_k | c = \varnothing\}, \tag{3}$$

and

$$B_k = \mathrm{Im}\, \partial_{k+1} = \{c \in C_k | \exists d \in C_{k+1} : c = \partial_{k+1} d\}. \tag{4}$$

where Z_k is the k-cycle group and B_k is the k-boundary group. With the aforementioned definitions, the k-homology group is defined to be the quotient group of the k-cycle group modulo the k-boundary group,

$$H_k = Z_k / B_k. \tag{5}$$

where Z_k is the k-homology group. The kth Betti number is defined to be rank of the k-homology group as $\beta_k = rank(H_k)$.

Filtration and persistent homology

Original homology is oversimplified for geometric analysis. Persistent homology introduces the nested sequence of subcomplexes to describe inclusive topological space which depends on a filtration parameter. Specifically, the filtration process of a simplicial complex K as a nested

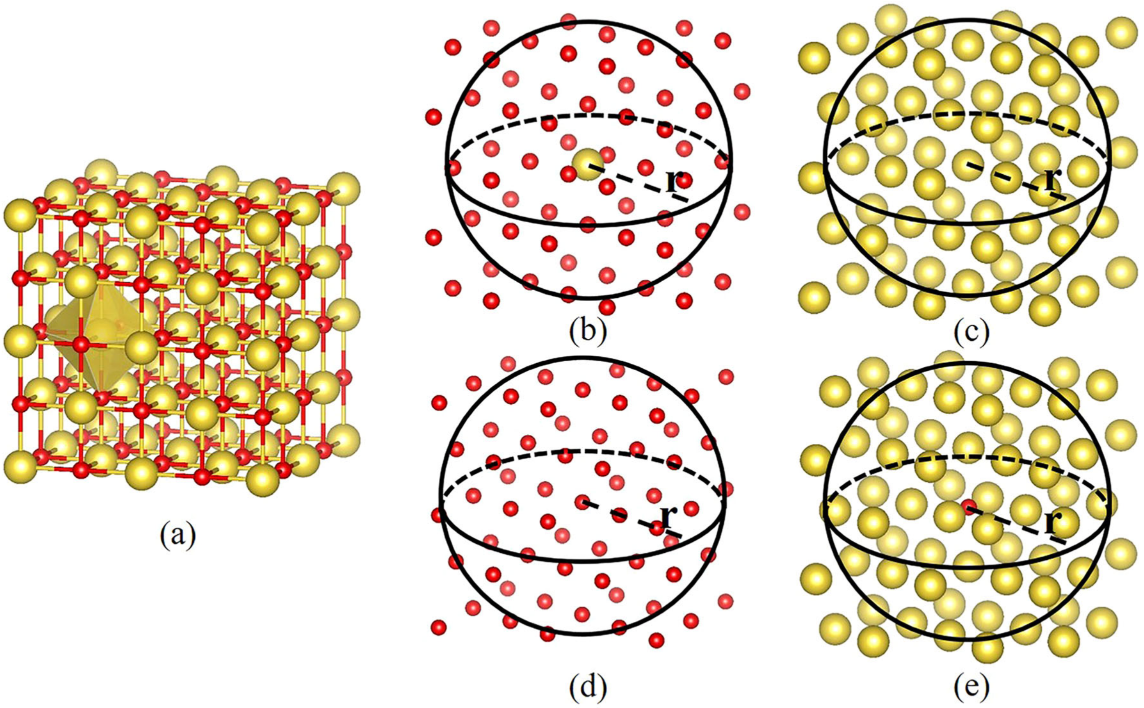

Fig. 5 Illustration of atom-specific persistent homology point clouds. **a** the original crystal structure of NaCl with red atom being Cl and yellow atom being Na. Four atom-specific point clouds are established by Na-Cl (**b**), Na-Na (**c**), Cl-Cl (**d**), and Cl-Na (**e**), respectively.

sequence of subcomplexes of K,

$$\varnothing \subseteq K_0 \subseteq K_1 \cdots \subseteq K_n = K. \tag{6}$$

Subcomplexes corresponding to various filtration parameters offer the topological fingerprints of multiple scales. The kth persistent Betti numbers $\beta_k^{i,j}$ are ranks of k_{th} homology groups of K_i that are alive and are defined as

$$\beta_k^{i,j} = \text{rank}(H_k^{i,j}) = \text{rank}\left(Z_k(K_i)/\left(B_k(K_j)\bigcap Z_k(K_i)\right)\right). \tag{7}$$

These persistent Betti numbers are used as topological fingerprints in machine learning studies of materials. There are different types of simplicial complex constructions used in persistent homology. The Vietoris–Rips (VR) complex used in this work is formed by all points in it has pairwise distances no greater than a cutoff distance d in a given metric space. The abstract property of the VR complex enables the construction of simplicial complexes for correlation function-based metric spaces, which models pairwise interaction of atoms with correlation functions instead of spatial metrics.

Atom-specific persistent homology

Persistent homology only offers the global structural information which cannot represent crystal structures with a wide range of chemical compositions and structural complexity. We introduce atom-specific persistent homology to embed atom-wise chemical information into topological invariants. The essential idea is that, in a unit cell, there are only a few atoms and each atom has its unique structural environment, which defines its own topological fingerprints.

Taking the classic NaCl crystal as an example, one can choose either the Na atom or the Cl atom as the atom of interest to generate atom-specific topological fingerprints. For each choice, there are two types of environments, namely Cl atoms or Na atoms. As a result, we have four possible combinations, namely Na-Na, Na-Cl, Cl-Na, and Cl-Cl. Their atom-specific point clouds are shown in Fig. 5.

In general, to capture element-level interactions, we consider the combination of all element pairs P for the substance composition. Given a specific composition, persistence barcodes are calculated as follows. The element-specific pair $P_{\alpha,i}^{\beta}$ represents a collection of pairs of atoms around the ith central atom of element type α and surrounding atoms of element type β, where α and β may be the same. First, expanding the unit cell so that the distance between the boundary atoms and any atoms in the original unit cell is smaller than the pre-defined cutoff radius r_c. Then, for the ith atom of interest in the unit cell, a point cloud consisting of all atoms within a cutoff radius r_c is selected

$$R_i^{\alpha,\beta}(r_c) = \{\mathbf{r}_j^{\beta} | \parallel \mathbf{r}_i^{\alpha} - \mathbf{r}_j^{\beta} \parallel < r_c, \mathbf{r}_j^{\beta}, \mathbf{r}_i^{\alpha} \in P_{\alpha,i}^{\beta}, \forall j \in 1, 2, \cdots, N\}, \tag{8}$$

where N is the number of atoms in pair $P_{\alpha,i}^{\beta}$. Given a point cloud, simplicial complex, homology group, and persistence barcode are computed via persistent homology. We compute the persistence barcodes by using software package Ripser[48]. The persistence barcode pair of central atom Na point cloud (Fig. 5b, c) is illustrated in Fig. 2. In the Betti-0 section of Na-Cl barcode, the six bars ended at 2.84 Å indicate that there are six nearest neighbor atoms Cl around central atom Na. The other bars ended at 4 Å show that the distance between any other two nearest neighbor atoms is 4 Å. There is no Betti-1 in this case because the distances between any two components are the same, which reflects the high-level symmetry of the structure.

Topological representations

The topological representations used in the machine-learning algorithms are extracted from persistence barcodes computed by atom-specific persistent homology. The cutoff radius r_c used to generate the barcodes in this paper is 12 Å. We describe the procedure for generating topological representations for crystalline compounds. The first step is to generate a collection of atom-specific barcodes denoted by $\{B(P_{\alpha,i}^{\beta}, D)\}$, where $P_{\alpha,i}^{\beta}$ was defined above, i goes through all atoms in the unit cell, α and β run over all possible element types, and D denotes topological dimensions, such as Betti-0, Betti-1, and Betti-2. Taking $BaTiO_3$ as an example, we will have Betti-0, Betti-1, and Betti-2 barcodes for each of five atoms in the unit cell. The second step is to generate a collection of element-specific barcodes denoted by $\{B(P_{\alpha}^{\beta}, D)\}$. This is done by combining together atom-specific barcodes according to their element types. Using $BaTiO_3$ as an example, we will have Betti-0, Betti-1, and Betti-2 barcodes for each of three element types. The third step is to characterize barcodes. In general, for any bar in one barcode, it is important to keep track of its birth, the death, and the persistent length, because this information is related to the bond length, ring, or cavity size. However, for Betti-0 bars, since their birth positions are uniformly 0, only the length of the bar needs to be recorded. The last step is to obtain statistics for each element type of barcodes. Therefore, for the element-specific barcodes of Betti-0 in $B(P_{\alpha}^{\beta}, D)$, five statistical quantities are calculated as the minimum, maximum, average,

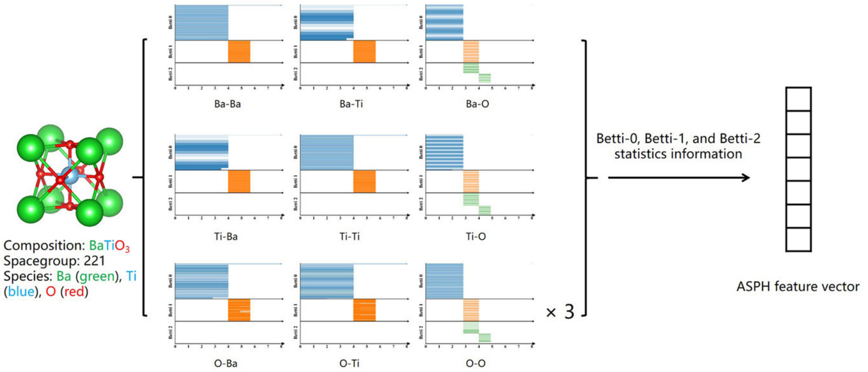

Fig. 6 The construction process of BaTiO$_3$ topological feature. First, compute all barcodes of the combinations formed by each atom in the unit cell as a central atom and surrounded by various element types. The ASPH features are then generated from the statistics data of these barcodes. Since the unit cell of BaTiO$_3$ having three O atoms, the barcode data of this type in the figure is multiplied by three.

standard error, and the sum of the bar length. Moreover, for element-specific barcodes of Betti-1 or Betti-2, we generate five statistical quantities, i.e., the minimum, maximum, average, standard deviation, and the sum for each of the birth, death, and persistent length. Therefore, we have a total of 35 element-specific topological representations for each element type. Additionally, we combine all atom-specific barcodes in the unit cell, which leads to another 35 statistical representations. For BaTiO$_3$, we $(3+1)\times 35$ (140) non-zero representations for BaTiO$_3$. Since there are 80 possible element types in the entire dataset, our total number of features is 2835 (i.e., 35×81). We set all representations to 0 for element types that do not exist in the molecule of interest. The overall process of element-specific representation generation is shown in Fig. 6. Basically, ASPH are translational, and rotational invariant by design and is able to reflect smooth changes due to perturbations in atomic positions (see Supplementary Figs. S3 and S4).

In addition to topological information described by ASPH, composition-based features are used in our method. These attributes are described in work by Ward et al.[49]. It contains the stoichiometric attributes for the fractions of element, elemental-property attributes based on statistics of the elemental properties of all atoms in the crystal, electronic structure attributes which are the average fraction of electrons from the *s*, *p*, *d* and *f* valence shells between all present elements[50], and ionic compound attributes consist of differences in electronegativities between constituent elements and whether it is possible to form an ionic compound if all elements in common oxidation states.

Machine learning algorithm, dataset, and validation

The ASPH and composition features are used as machine learning features to predict inorganic periodic solids formation enthalpies. For ML algorithm selection, we choose to use gradient boosted regression trees (GBRT)[51] to test the accuracy, robustness, and efficiency of topological based features. GBRT is able to combine a number of weak predictors to create a strong model. The training of a GBRT model is done by adding one tree at a time to reduce the lose function of the current model. In practice, different randomly selected subsets of the training data and features are used for each update of the model to reduce overfitting. Hyper-parameter searching is done by the cross-validation judged by R^2. The hyper-parameters used in GBRT are: n_estimators = 300,000, learning_rate = 0.001, max_depth = 7, min_samples_split = 5, subsample = 0.85 and max_features = sqrt. The ML models are built using scikit-learn software (version 0.19.2)[52]. Our dataset includes 31912 compounds which primitive cell size smaller than 40 atoms, covering the seven lattice systems and 80 elements (H-Pu, excluding noble gases, Tc, Pa, Pm, Po, At, Rn, Fr, Ra, and Ac). Tenfold cross-validation of the data sets is used to verify the proposed method. To address the robustness of the machine learning model, the random splitting of data in tenfold cross-validation is repeated 20 times. The median performance and the standard deviation of the performance across repeated experiments are reported. The replication of Voronoi tessellations and Coulomb Matrix are using Magpie, which is freely available under an open-source license[49].

DATA AVAILABILITY

The experimental data in this work is in the same Github repository as the code. The structure data is obtained from ICSD[47], and the DFT-calculated property data is from OQMD[7].

CODE AVAILABILITY

The code used to generate results in the manuscript is available under the Github repo: https://github.com/PKUsamPHTeam/ASPH-Code.

Received: 22 June 2020; Accepted: 6 January 2021;
Published online: 05 February 2021

REFERENCES

1. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
2. Isayev, O. et al. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 1–12 (2017).
3. Walsh, A. The quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
4. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev. Mater.* **136**, B864 (1964).
5. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
6. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *Appl Mater.* **1**, 011002 (2013).
7. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *JOM* **65**, 1501–1509 (2013).
8. Curtarolo, S. et al. Aflowlib. org: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
9. Jie, J. et al. A new materialgo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps. *Sci. Chin. Technol. Sci.* **62**, 1423–1430 (2019).
10. Sutton, R. S. et al. Introduction to reinforcement learning, **135** (MIT press Cambridge, 1998).
11. Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC$_2$D$_6$) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
12. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
13. Kim, K. et al. Machine-learning-accelerated high-throughput materials screening: discovery of novel quaternary heusler compounds. *Phys. Rev. Mater.* **2**, 123801 (2018).
14. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
15. Lee, J., Seko, A., Shitara, K., Nakayama, K. & Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
16. Rajan, A. C. et al. Machine-learning-assisted accurate band gap predictions of functionalized mxene. *Chem. Mater.* **30**, 4031–4038 (2018).

17. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
18. Lu, S. et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 1–8 (2018).
19. Jie, J. et al. Discovering unusual structures from exception using big data and machine learning techniques. *Sci. Bull.* **64**, 612–616 (2019).
20. Seko, A. et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. *Phys. Rev. Lett.* **115**, 205901 (2015).
21. Seko, A., Hayashi, H., Nakayama, K., Takahashi, A. & Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **95**, 144110 (2017).
22. Sosso, G. C., Deringer, V. L., Elliott, S. R. & Csányi, G. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials. *Mol. Simul.* **44**, 866–880 (2018).
23. Furmanchuk, A., Agrawal, A. & Choudhary, A. Predictive analytics for crystalline materials: bulk modulus. *RSC Adv.* **6**, 95246–95251 (2016).
24. Evans, J. D. & Coudert, F.-X. Predicting the mechanical properties of zeolite frameworks by machine learning. *Chem. Mater.* **29**, 7833–7839 (2017).
25. Lu, Y., Chen, X., Zhao, C.-Z. & Zhang, Q. Machine learning towards screening solid-state lithium ion conductors. *Chin. J. Struct. Chem.* **1**, 2 (2020).
26. Takahashi, A., Seko, A. & Tanaka, I. Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: application to elemental titanium. *Phys. Rev. Mater.* **1**, 063801 (2017).
27. Hu, Q. et al. Neural network force fields for metal growth based on energy decompositions. *J. Phys. Chem. Lett* **11**, 1364–1369 (2020).
28. Butler, K. T., Frost, J. M., Skelton, J. M., Svane, K. L. & Walsh, A. Computational materials design of crystalline solids. *Chem. Soc. Rev.* **45**, 6138–6146 (2016).
29. Shi, S. et al. Multi-scale computation methods: their applications in lithium-ion battery research and development. *Chin. Phys. B* **25**, 018212 (2015).
30. Weng, M. et al. Identify crystal structures by a new paradigm based on graph theory for building materials big data. *Sci. Chin. Chem.* **62**, 982–986 (2019).
31. Ulissi, Z. W., Medford, A. J., Bligaard, T. & Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and dft calculations. *Nat. Commun.* **8**, 1–7 (2017).
32. Nguyen, D. D., Cang, Z. & Wei, G.-W. A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **22**, 4343–4367 (2020).
33. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
34. Braams, B. J. & Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **28**, 577–606 (2009).
35. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
36. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
37. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem.* **115**, 1094–1101 (2015).
38. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
39. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
40. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
41. Kaczynski, T., Mischaikow, K. & Mrozek, M. *Computational homology*, vol. 157 (Springer Science & Business Media, 2006).
42. Wu, K., Zhao, Z., Wang, R. & Wei, G.-W. Topp–s: Persistent homology-based multi-task deep neural networks for simultaneous predictions of partition coefficient and aqueous solubility. *J. Comput. Chem.* **39**, 1444–1454 (2018).
43. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geomet* **33**, 249–274 (2005).
44. Cang, Z. & Wei, G.-W. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. *Int. J. Numer. Method. Biomed. Eng.* **34**, e2914 (2018).
45. Cang, Z. & Wei, G.-W. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics* **33**, 3549–3557 (2017).
46. Cang, Z. & Wei, G.-W. Topologynet: topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS Comput. Biol.* **13**, e1005690 (2017).
47. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallogr Section B: Struct Sci* **58**, 364–369 (2002).
48. Tralie, C., Saul, N. & Bar-On, R. Ripser. py: a lean persistent homology library for python. *J. Open Source Softw.* **3**, 925 (2018).
49. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 16028 (2016).
50. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
51. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
52. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

ACKNOWLEDGEMENTS

The research was financially supported by Soft Science Research Project of Guangdong Province (2017B030301013), National Key R&D Program of China (2016YFB0700600), and Shenzhen Science and Technology Research Grant (ZDSYS201707281026184). The work of Guo-Wei Wei was supported in partial by NSF Grants DMS1721024, DMS1761320, IIS1900473, NIH grants GM126189 and GM129004, Bristol-Myers Squibb, and Pfizer.

AUTHOR CONTRIBUTIONS

Y.J. designed the project and carried out the experiment. Y.J., D.C., X.C., T.L., G.-W.W., and F.P. discussed the results, analyzed the data, and drafted the manuscript. G.-W.W. and F.P. conceptualized the project and obtained funding.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00493-w.

Correspondence and requests for materials should be addressed to G.-W.W. or F.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.