Persistent homology-based descriptor for machine-learning potential

Emi Minamitani$^{1,2,3}$ and Ippei Obayashi$^{3,4}$

1. The Institute of Scientific and Industrial Research, Osaka University, Ibaraki 567-0047, Japan
2. Institute for Molecular Science, Okazaki 444-8585, Japan
3. JST, PRESTO, Kawaguchi, Saitama 332-0012, Japan
4. Center for Artificial Intelligence and Mathematical Data Science, Okayama University, Okayama 700-8530, Japan

*E-mail: eminamitani@sanken.osaka-u.ac.jp
Abstract

Constructing efficient descriptors for representing atomic configurations is crucial for developing superior machine-learning potentials. The widely used conventional descriptors are local and constructed based on the two- and three-body correlations of atomic distribution around a specific atom. This atom-centered approach is convenient for constructing a machine-learning potential with scalability and invariance under symmetry operations. However, these descriptors show several limitations with respect to the classification of different configurations, which have a detrimental effect on the predictions of physical properties. Here, we show that the persistent diagram, that is, the two-dimensional representation of persistent homology, can be used as a descriptor. First, using the clustering of cyclo-octane conformations as an example, we show that the persistent diagram captures both the local geometrical and global topological characteristics of the atomic configuration. Next, we demonstrate that convolutional neural network models based on the persistent diagram can accurately predict the mean energies per atom of amorphous graphene and amorphous carbon. Moreover, the models can predict the energies for systems larger than those used in the training process, meaning that they are scalable with respect to the system size. Our results provide an effective strategy for improving the machine-learning potential using descriptors that depict both geometrical and topological information.

Keywords: topological data analysis, persistent homology, machine-learning potential, amorphous, convolutional neural network
1. Introduction

Using machine-learning potential is a promising approach for resolving the trade-off between computational accuracy and the cost of predicting the physical properties of complex systems. Various machine-learning models have been developed over the past decade, such as Gaussian process regression\textsuperscript{1,2} and artificial neural networks\textsuperscript{3–6} and their derivatives\textsuperscript{7–12}. The properties of the model structure used for each technique are different. However, the basic strategy is the same: constructing a surrogate model of the relationship between the atomic coordinates and physical properties, which can be used as an alternative to computationally demanding calculations.

Developing a suitable method to convert the atomic coordinates into vectorial data is crucial for this purpose. Various methods have been proposed to obtain vectorial representations of the atomic coordinates, often called descriptors\textsuperscript{13}. These descriptors can be classified into two types: local and global descriptors. Local descriptors are constructed using atom-centered approaches, wherein the many-body correlations of the atomic distributions around each atom are regarded as the local chemical environment\textsuperscript{14}. Examples of such descriptors are atom-centered symmetry functions (ACSFs)\textsuperscript{3,15} and smooth overlap atomic positions (SOAPs)\textsuperscript{16,17}. One advantage of developing the machine-learning potential using local descriptors is scalability: one can predict the properties of a larger system than that used for training. In addition, most local descriptors naturally fulfill the physical requirements, that is, they exhibit invariance under translation, rotation, and reflection. The atom-centered approach also makes it possible to ensure invariance with respect to permutations of chemically equivalent atoms by using a simple model architecture\textsuperscript{3,15–17}. However, the local descriptors have a fundamental
limitation, as demonstrated by Pozdnyakov et al.\textsuperscript{18}. The two- and three-body correlations encoded in the local descriptors are not complete, and this can lower the prediction accuracy. Moreover, many hyperparameters, such as the cutoff, shape, and size of the basis functions are necessary for constructing the local descriptors. The optimal hyperparameters vary depending on the system, and their rigorous optimization, which is required for the practical use of these machine-learning potentials, is both time-consuming and difficult.

In contrast, global descriptors such as the Coulomb matrix\textsuperscript{19} and bag-of-bonds\textsuperscript{20} have fewer hyperparameters and may contain information related to higher-order many-body correlations. However, global descriptors are not readily scalable because of the rapid increase in the computational cost with the number of atoms. Additionally, they require specific treatments or complex model architectures to maintain invariance with respect to the permutations of chemically equivalent atoms. Thus, an alternative to conventional local and global descriptors is required.

In this study, we focus on persistent homology (PH), an emerging technique in computational topology, which is superior for extracting shape characteristics\textsuperscript{21–23}. This technique has been widely used for the structural analysis of various types of condensed matters ranging from biological molecules\textsuperscript{24} to glass\textsuperscript{25–30}. Recently, it has been demonstrated that PH is also useful for determining the relationship between the physical properties of materials and their atomic structures\textsuperscript{31–37}.

Let us briefly introduce the basics of the PH. In mathematics, a shape is considered as a topological space, $X$, consisting of a set of points (in our case, atoms) and a topology.
The topological characteristics of $X$ are represented by holes in space, which are encoded in an algebraic structure called the $k$-th homology group, $H_k(X)$. A nested sequence of topological space, $X_1 \subseteq X_2 \subseteq \cdots \subseteq X_n$, results in the series $H_k(X_1), H_k(X_2), \cdots, H_k(X_n)$. The idea of PH involves tracking the elements of $H_k(X_i)$ as $i$ increases: here $i$ represents the scale and is often referred to as “time.” The procedure used to obtain such a sequence is called filtration, which is schematically shown in Fig. 1. In the filtration process, for the first homology in a two-dimensional system, circles with radius $r$ are placed at the respective atoms. Subsequently, $r$ is gradually increased. With increasing $r$, the circles start to intersect, and we set edges between the centers of the circles. The growing network formed by these edges defines the sequence of the topological space, $X_i$. A closed path formed by these edges corresponds to a topological feature called a cycle. With a further increase in the radius, the closed path is fully covered by these circles. This is interpreted as a cycle being converted into another class of topological features called a boundary. The features of PH are represented by the pairs of birth and death times at which the cycles appear and are converted into boundaries. Filtration can be generalized into three or more dimensions using spheres or hyperspheres. The two-dimensional visualization of the pairs of birth and death times is called a persistence diagram (PD).

As per the above definition, the PD of an atomic structure is invariant under the translation, rotation, reflection, and permutation of the same kind of atoms. The PD is expected to hold information regarding both the connectivity and atomic distribution, which is automatically collected during the construction of the PH without manual feature engineering. In addition, there is no restriction on the order of the correlations during the
construction of the PD, and the PD can include information on higher-order correlations. These characteristics suggest that the PD can be used as a descriptor.

In this study, we propose a new type of descriptor constructed from the PD and use it to predict the energy accurately based on a machine-learning potential. To this end, we first compare the PD descriptor with the other conventional descriptors using a dataset of cyclo-octane conformations. We visualize the differences in the information captured by the respective descriptors by dimensional reduction using principal component analysis (PCA). The three canonical conformations of cyclo-octane are adequately clustered when the global information and permutation invariance are both maintained. The PD descriptors fulfill these conditions and provide the proper clustering in the projected low-dimensional space. Next, we show that a convolutional neural network (CNN) model based on the PD descriptor exhibits a sufficiently high accuracy for predicting the mean energy per atom of amorphous graphene (aGr) and amorphous carbon (aC). For aC, the root mean square error (RMSE) for the mean energy per atom was 27.7 meV/atom for the training data and 33.3 meV/atom for the test data; these values are comparable to those obtained in previous studies. Thus, our study illustrates an alternative for describing atomic configurations based on topological information to improve machine-learning potential.

2. Methods

2.1 PH calculation

The HomCloud code \(^{38}\) was used to analyze the PH of the amorphous models. In this study, we focused on the first homology.
We constructed the PD descriptors in the following manner. For the analysis of the cyclo-octane dataset, the PDs were converted into vectors using the persistent image method\textsuperscript{39,40}. The persistent image corresponds to the distribution of the birth–death pairs, and is defined as follows:

\[
\rho(x_i, y_i) = \sum_k w(b_k, d_k) \exp\left( -\frac{(b_k - x_i)^2 + (d_k - y_i)^2}{2\sigma^2} \right) \tag{1}
\]

where \((b_k, d_k)\) is the k-th pair of the birth and death times. Note that we used the squared values of the radii of the spheres employed in the filtration procedure as the birth and death times. Therefore, their unit was Å\(^2\). We set the weight, \(w(b_k, d_k)\), as a constant at 1.0. The standard deviation of the Gaussian, \(\sigma\), was set to 0.05. We focused on the square region \([0.5, 3.3]_{\text{birth}} \times [1.95, 3.3]_{\text{death}}\), which was divided into 200 \times 100 meshes.

We evaluated \(\rho(x_i, y_i)\) on these meshes and reshaped the results into a vector in \(\mathbb{R}^{15278}\).

For the input of the machine-learning potential of aGr and aC, we converted the PDs into two-dimensional histograms with a 128\times 128 mesh. We focused on the square region \([0.0, 4.0]_{\text{birth}} \times [0.0, 4.0]_{\text{death}}\) for aGr, and \([0.0, 8.0]_{\text{birth}} \times [0.0, 8.0]_{\text{death}}\) for aC. As the PD reflects the number and size of the ring structures, the histogram values depend on the system size. Therefore, we converted the histogram into a probability distribution and standardized it to make the descriptors independent of the model size. We used the max-min standardization for aGr and the z-score standardization for aC. A periodic boundary condition was imposed while calculating the PDs for aC because of the small model size.

2.2 Dataset generation
For aGr, the amorphous structures of 680 atoms were determined via melt-quench simulations based on classical molecular dynamics (MD) using LAMMPS. The dataset contained the atomic coordinates and total energies of 1,830 structures: 1,050 were sampled from the final state of the melt-quench simulations at different cooling rates, and 780 were sampled from the intermediate state during the simulations. For aC, the samples were prepared via ab-initio MD melt-quench simulations using VASP. Cubic cells of different sizes containing 216 atoms were simulated to produce amorphous models with densities of 2.0–3.4 g/cm$^3$. We performed 48 independent MD runs and sampled intermediate structures from the trajectories of these runs. This resulted in a dataset of atomic coordinates and total energies for 12,960 structures. The detailed conditions used for these simulations are provided in the Supplementary Materials Sections 1 and 2.

3. Results and discussion
3.1 Representation of the cyclo-octane conformations

First, we determined the differences in the structural representations in the PDs and those in conventional descriptors. We choose the cyclo-octane dataset previously used in Refs. for this purpose. The dataset consists of the C-atom positions in the cyclo-octane molecules with 6,040 conformations. These were generated by introducing various torsion angles for a fixed bond angle ($115^\circ$) and bond length (1.52 Å). Therefore, simple two- and three-body correlations between the nearest neighbors cannot distinguish the conformations in this dataset. As shown in Figs. 2a and b, the PDs for the two conformations selected from the dataset are visibly different, indicating that the PDs include information for higher-order many-body correlations.
We visualized the differences between the conventional descriptors and the PD descriptors based on the dimensional reduction using PCA. The conformations in the dataset are the combinations of three canonical conformations called chair, boat, and crown, which are shown in Fig. 3a. Thus, whether these three conformations are properly clustered in the projection to the low-dimensional space or not would be a reasonable measure to assess the efficiency of the descriptors.

The simplest descriptor of atomic positions is the set of Cartesian coordinates. The Cartesian coordinates of the eight C atoms in cyclo-octane are points in $\mathbb{R}^{24}$. In the projection $\mathbb{R}^{24} \rightarrow \mathbb{R}^2$ obtained using PCA, the chair and crown conformations are not separated. Moreover, while the boat conformations are separated from the other conformations, they are split into four groups (Fig. 3b).

Figures 4a and b exhibit the PCA results for the high-dimensional vector representations from the local descriptors. We focused on a single C atom for the respective cyclo-octane conformations and constructed SOAP and ACSF descriptors centered at the atom. Note that the choice of the target atom does not affect the calculation results, as shown in the Supplementary Materials Section 3. The hyperparameters used to construct the descriptors are also listed in the same section. In the projection of these local descriptors on $\mathbb{R}^2$, the chair, boat, and crown conformations are separated. However, the chair conformations are split into two groups, which indicates that the similar conformations may be assigned to spatially separated points in the local descriptor space. When we use the average of the SOAP descriptors obtained for eight C atoms to include the global information, the chair, boat, and crown conformations are adequately clustered, as shown
in Fig. 4c. In compensation for the better clustering, many conformations are assigned to the same point, meaning that using the average cancels the minute differences between the conformations.

Next, we evaluated the dimensional reduction of the global descriptors. As a representative global descriptor, we evaluated the Coulomb matrix. In the projection $\mathbb{R}^{64} \rightarrow \mathbb{R}^2$, the boat and crown conformations are not separated, and the chair conformations are split into four groups (Fig. 4d). One reason of the poor clustering performance, even though the global information was included, is the lack of invariance under the permutation of the atoms. By using the sorted eigenvalues of the Coulomb matrix, we could obtain the global descriptor with invariance under the permutation. The dimensional reduction of this eigen spectrum descriptor from $\mathbb{R}^8$ to $\mathbb{R}^2$ yields results similar to those seen in the case of the averaged SOAP descriptors (Fig. 4e).

The above-described results for the local and global descriptors highlight the importance of both the global information and invariance under permutation during embedding of the atomic coordinates in the descriptor space. The PD descriptors satisfy these conditions and indeed result in proper clustering, as shown in Fig. 4f. This result suggests that the PDs embed the atomic coordinates appropriately: the similar structures are close to each other, while the different ones are suitably separated. It should be mentioned that the PD descriptors sometimes failed to distinguish between the different conformations of cyclo-octane. One reason for this is the filtration method used to construct the PDs. We used alpha complexes, in which the area covered by a sphere was restricted by the Voronoi diagram during filtration. While this treatment reduces the computational cost of
calculating the PDs, it also limits the amount of information on a long scale. In the case of the cyclo-octane dataset, the combination of the above-mentioned condition and the fixed bond lengths and angles sometimes cause coincidence of the birth and death times. We found that the longer scale information can be enhanced by using the PDs evaluated on the virtual molecular structure where one C atom was removed from the original structure. Concrete examples are given in Supplementary Materials Section 3. The above-described coincidence is less likely to occur in the case of complex systems such as amorphous.

3.2 Application of PDs to machine-learning potential for amorphous solids

Based on the potential of PD descriptors discussed in Sec. 3.1, we applied the PDs as the inputs for the machine-learning potential to predict the mean energies per atom for aGr and aC. We chose these systems as representatives of two- and three-dimensional materials, respectively. Examples of the structures in the aGr and aC datasets are shown in Figs. 5a and b, respectively. The PDs of these structures were converted into normalized two-dimensional histograms, as shown in Figs. 5c and d. These histograms have the same format as grayscale images with a pixel size of 128 × 128. These PDs encode the structural characteristics as the distribution of the birth and death times, that is, how they appear as images. Based on this consideration, we constructed a learning model based on a CNN architecture that exhibited excellent image recognition performance. A schematic of the model is shown in Fig. 5e. The details of the model is described in the Supporting Information Section 4.1.

The results of the training and testing of the learning model on the datasets of aGr and
aC are shown in Figs. 6a and b, respectively. For aGr (aC), the RMSE for the mean energy per atom was 7.9 meV/atom (27.7 meV/atom) for the training data and 8.3 meV/atom (33.3 meV/atom) for the test data. The difference in the RMSE values of the training and test data was small, and the learning curve gradually decreased for both the test and training data (Supporting Information Section 4.1), indicating that overfitting was suppressed.

A prerequisite for the machine-learning potential is to ensure that the accuracy of the predictions is high when the potential is used to a larger-size system. Therefore, we predicted the mean energies per atom in aGr (aC) with 2,720 (512) atoms. The data for the larger amorphous structure and energy were generated using the same method as that employed with the original data set. The prediction results are shown in Figs. 6c and d. The RMSE for the 2,720-atom aGr system was 3.1 meV/atom, and that for the 512-atom aC system was 33.6 meV/atom. These results indicate that the obtained machine-learning models fulfill the requirements.

In the case of aC, the samples with a mean energy per atom of more than -8.0 eV correspond to the liquid phase that appeared during the melt-quench simulation. The liquid phase does not exhibit any specific topological features such as medium-range order; nevertheless, the accuracy of the energy predictions based on the topological descriptors remained high. This could be because the PDs retain not only the pure topological information (how the atoms are connected) but also the geometric information (the distance and angle between the atoms and their correlations). We also evaluated the ability of the model to interpolate. As summarized in Supporting Information Section 4.2,
the model is robust against the missing data in the middle-energy region, which indicates high interpolation ability.

Furthermore, we visualized the activation mapping obtained from the outputs and gradients of the final convolution layer to elucidate the basis of the prediction (for details, see Supporting Information Section 4.3). Such explainability for model predictions is one of the superior points of PD descriptor. As shown in Fig. 7a, the learning model predicted the energy in aC for a high-energy liquid-like system based on the distribution of the birth–death pairs around the diagonal line. On the other hand, the distribution of the birth–death pairs away from the diagonal line was focused on the low-energy amorphous sample (Fig. 7b). This flexibility with respect to the attention region resulted in accurate predictions.

The RMSE of the energy prediction for the aC dataset, which is 33.3 meV/atom, is comparable to that obtained using the Gaussian approximation potential (GAP) as well as those for the high-dimensional neural network potential (HDNNP) for carbon reported in previous studies. For a more quantitative comparison, we prepared the local descriptors for the same 216-atom aC dataset using SOAP and trained an HDNNP model to predict the mean energy per atom. As shown in Fig. 8a, the accuracy of the prediction based on the HDNNP model using the SOAP descriptor was similar to that for the CNN model using the PD descriptor. We would like to emphasize that, in the former case, the prediction accuracy depends on the hyperparameters used for the SOAP construction. As shown in Fig. 8b, when we used SOAP descriptors constructed using slightly different hyperparameters, the RMSE of the prediction for the test data became larger than 100
meV. This highlights the need to fine tune the hyperparameters. In contrast, PD descriptors are almost free from hyperparameters, and can achieve high accuracy without requiring a tuning process.

4. Conclusion
In summary, we constructed a new type of descriptor for atomic structures using PH. We propose that a PD captures both the local geometrical and global topological characteristics of atomic configurations and describes the structural characteristics differently from conventional descriptors. As a demonstration, we showed that the different conformations of cyclo-octane molecules can be appropriately classified in the low-dimensional space obtained from the PCA for PD descriptors. We also demonstrated that a CNN model based on a PD descriptor exhibited sufficient accuracy with respect to the prediction of the mean energy per atom for amorphous structures. The strategy of using PDs as the descriptors of atomic structures is different from using the conventional descriptors based on the correlation of atomic distribution and is expected to improve the reliability of machine-learning models for predicting the local and global physical properties. An area of future work will be the applicability of this PD descriptor for the MD simulation in large-scale systems. It is also worth investigating whether PD descriptors can be used along with the ACSFs and SOAPs as the inputs for the HDNNP and GAP. Exploring these possibilities should improve the machine-learning potential by using descriptors that depict both the topological and geometric information.
Associated content
Supplementary Materials
Computational setup for generating amorphous graphene dataset; computational setup for generating amorphous carbon dataset; analyses of cyclo-octane dataset; machine-learning model

Author information
Corresponding Author
Emi Minamitani, The Institute of Scientific and Industrial Research, Osaka University, Ibaraki 567-0047, Japan
Email: eminamitani@sanken.osaka-u.ac.jp
ORCID: 0000-0002-8003-6526

Author
Ippei Obayashi, Cyber-Physical Engineering Informatics Research Core (Cypher), Okayama University, Okayama 700-8530, Japan
ORCID: 0000-0002-7207-7280

Notes
The authors declare no competing financial interests.

Acknowledgements
This study was supported by PRESTO Grants (Grant Numbers JPMJPR1923, and JPMJPR2198), JST, and KAKENHI Grants (Grant Numbers 21H01816, 19H02544, 19H00834, and 20H05884), MEXT, Japan. The calculations were performed using a computer facility at the Research Center for Computational Science (Okazaki, Japan).

References
1 W.J. Szlachta, A.P. Bartók, and G. Csányi, Phys. Rev. B Condens. Matter 90, 104108 (2014).
2 A.P. Bartók, M.C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. 104, 136403 (2010).
3 J. Behler and M. Parrinello, Phys. Rev. Lett. 98, 146401 (2007).
4 J.S. Smith, O. Isayev, and A.E. Roitberg, Chem. Sci. 8, 3192 (2017).
5 K.T. Schütt, P.-J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, (2017).
6 J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, and G.E. Dahl, (2017).
7 A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker, J. Comput. Phys. 285, 316 (2015).
8 V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C 121, 511 (2017).
9 A.V. Shapeev, Multiscale Model. Simul. 14, 1153 (2016).
10 L. Zhang, J. Han, H. Wang, R. Car, and W. E, Phys. Rev. Lett. 120, 143001 (2018).
11 J. Gasteiger, J. Groß, and S. Günemann, arXiv:2003.03123 [cs.LG] (2020).
12 O.T. Unke and M. Meuwly, J. Chem. Theory Comput. 15, 3678 (2019).
13 M. Pinheiro, F. Ge, N. Ferré, P.O. Dral, and M. Barbatti, Chem. Sci. 12, 14396 (2021).
14 M.J. Willatt, F. Musil, and M. Ceriotti, J. Chem. Phys. 150, 154110 (2019).
15 J. Behler, J. Chem. Phys. 134, 074106 (2011).
16 S. De, A.P. Bartók, G. Csányi, and M. Ceriotti, Phys. Chem. Chem. Phys. 18, 13754 (2016).
17 A.P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B Condens. Matter 87, 184115 (2013).
18 S.N. Pozdnyakov, M.J. Willatt, A.P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. 125, 166001 (2020).
19 M. Rupp, A. Tkatchenko, K.-R. Müller, and O.A. von Lilienfeld, Phys. Rev. Lett. 108, 58301 (2012).
20 K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, J. Phys. Chem. Lett. 6, 2326 (2015).
21 A. Zomorodian and G. Carlsson, Proceedings of the Annual Symposium on Computational Geometry 274, 347 (2004).
22 H. Edelsbrunner and J. Harer, Computational Topology: An Introduction (Amer Mathematical Society, 2010).
23 M. Buchet, Y. Hiraoka, and I. Oabayashi, in Nanoinformatics (Springer Singapore, 2018), pp. 75–95.
24 K. Xia and G.-W. Wei, Int. j. Numer. Method. Biomed. Eng. 30, 814 (2014).
25 Y. Hiraoka, T. Nakamura, A. Hirata, E.G. Escolar, K. Matsue, and Y. Nishiura, Proc. Natl. Acad. Sci. U. S. A. 113, 7035 (2016).
26 T. Nakamura, Y. Hiraoka, A. Hirata, E.G. Escolar, and Y. Nishiura, Nanotechnology 26, 304001 (2015).
27 Y. Onodera, S. Kohara, P.S. Salmon, A. Hirata, N. Nishiyama, S. Kitani, A. Zeidler, M. Shiga, A. Masuno, H. Inoue, S. Tahara, A. Polidori, H.E. Fischer, T. Mori, S. Kojima, H.
Kawaji, A.I. Kolesnikov, M.B. Stone, M.G. Tucker, M.T. McDonnell, A.C. Hannon, Y. Hiraoka, I. Obayashi, T. Nakamura, J. Akola, Y. Fujii, K. Ohara, T. Taniguchi, and O. Sakata, NPG Asia Materials 12, 85 (2020).

28 M. Murakami, S. Kohara, N. Kitamura, J. Akola, H. Inoue, A. Hirata, Y. Hiraoka, Y. Onodera, I. Obayashi, J. Kalikka, N. Hirao, T. Musso, A.S. Foster, Y. Idemoto, O. Sakata, and Y. Ohishi, Phys. Rev. B Condens. Matter 99, 45153 (2019).

29 S. Hong and D. Kim, J. Phys. Condens. Matter 31, 455403 (2019).

30 S.S. Sørensen, C.A.N. Biscio, M. Bauchy, L. Fajstrup, and M.M. Smedskjaer, Science Adv. 6, eabc2320 (2020).

31 Y. Jiang, D. Chen, X. Chen, T. Li, G.W. Wei, and F. Pan, Npj Computational Materials 7, (2021).

32 J. Townsend, C.P. Micucci, J.H. Hymel, V. Maroulas, and K.D. Vogiatzis, Nat. Commun. 11, 1 (2020).

33 G. Kusano, K. Fukumizu, and Y. Hiraoka, in 33rd International Conference on Machine Learning, ICML 2016 (2016), pp. 2948–2957.

34 E. Minamitani, T. Shiga, M. Kashiwagi, and I. Obayashi, J. Chem. Phys. 156, 244502 (2022).

35 X. Chen, D. Chen, M. Weng, Y. Jiang, G.W. Wei, and F. Pan, J. Phys. Chem. Lett. 11, 4392 (2020).

36 S.M. Moosavi, H. Xu, L. Chen, A.I. Cooper, and B. Smit, Chem. Sci. 11, 5423 (2020).

37 Y. Lee, S.D. Barthel, P. Dłotko, S.M. Moosavi, K. Hess, and B. Smit, Nat. Commun. 8, 15396 (2017).

38 HomCloud, (https://homcloud.dev/index.en.html).

39 H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, J. Mach. Learn. Res. 18, 1 (2017).

40 I. Obayashi, Y. Hiraoka, and M. Kimura, Journal of Applied and Computational Topology 1, 421 (2018).

41 S. Plimpton, J. Comput. Phys. 117, 1 (1995).

42 G. Kresse and J. Hafner, Phys. Rev. B Condens. Matter 47, 558 (1993).

43 G. Kresse and J. Hafner, Phys. Rev. B Condens. Matter 49, 14251 (1994).

44 G. Kresse and J. Furthmüller, Phys. Rev. B Condens. Matter 54, 11169 (1996).

45 G. Kresse and J. Furthmüller, Comput. Mater. Sci. 6, 15 (1996).

46 S. Martin, A. Thompson, E.A. Coutsias, and J.P. Watson, J. Chem. Phys. 132, 234115 (2010).

47 B.J. Stolz, J. Tanner, H.A. Harrington, and V. Nanda, Proc. Natl. Acad. Sci. U. S. A.
Figure captions

Figure 1
Schematics of the filtration procedure used to obtain PD from data points. In this example, there are five points (a, b, c, d, and e). The PH group obtained by filtration is represented by two birth–death pairs, A and B, in the PD. Values of birth and death time are 0.96 and 1.02 at A and 1.05 and 1.30 at B. Pair A (B) corresponds to the birth and death of the cycle defined by the closed path a-b-c-a (c-b-d-e-c). The grey-shaded polygon in growing sequence of topological space indicates that the closed path formed by the edges of the polygon is fully covered by circles during filtration process.

Figure 2
(a) and (b) PDs of two conformations in the cyclo-octane dataset. Insets are the structural models of respective conformations.

Figure 3
(a) Schematics of three canonical conformations named chair, boat, and crown. (b) Projections of Cartesian coordinates of eight C atoms on $\mathbb{R}^2$ as determined by dimensional reduction using PCA. Horizontal and vertical axes represent values of 1st and 2nd principal components, respectively. The blue, red, and green points correspond to the conformations assigned to chair, boat, and crown, respectively.

Figure 4
Projections of the high-dimensional descriptors on $\mathbb{R}^2$ as determined by dimensional reduction using PCA. (a) and (b) Results obtained using SOAP and ACSF descriptors.
centered at C atom in cyclo-octane. These descriptors were constructed using the DScribe package$^{47}$, and can be considered points in $\mathbb{R}^{324}$ and $\mathbb{R}^{44}$, respectively. (c) Results obtained using the average of the SOAP descriptors at eight C atoms in cyclo-octane. (d) Results for the Coulomb matrix case. (e) Results for the the descriptor obtained from the sorted eigenvalue of the Coulomb matrix. (f) Projection of the descriptors in $\mathbb{R}^{15278}$ constructed from PDs using the persistent image$^{34,35}$. The blue, green, and red points in (a)–(f) correspond to conformations assigned to chair, boat, and crown, respectively. The details of the parameters used in SOAP, ACSF, and the persistent image descriptors are described in Supporting Information Section 1.

Figure 5
Examples of the amorphous structures of (a) graphene (aGr) and (b) carbon (aC). (c) and (d) Descriptors constructed from the PDs for the structures shown in (a) and (b), respectively. Differences in the distributions and values for these two descriptors are attributed to the difference in the focus areas and standardization methods. e) Machine learning model used in this study. The model consists of three convolution layers (conv) and two fully connected layers (FC). The max pooling layer (Max Pool) for down-sampling is placed after each convolution layer. The size of the filter is $3 \times 3$ in all the convolution layers. The first two convolution layers have 64 channels, and the last one has 32 channels. Therefore, the input size of the initial FC layer is $(128/8) \times (128/8) \times 32=8192$.

Figure 6
Comparison of the calculation results and predictions by the machine-learning model: mean energies per atom in (a) aGr and (b) aC. The red (blue) dots represent comparison results for the test (training) data. The predictions of mean energies per atom for larger systems using the same model for aGr and aC are shown in (c) and (d). The parameters used for machine learning are given in Supporting Information Section 4.1.

Figure 7
(a) and (b) Activation maps for high- and low-energy samples for the predictions of the aC dataset with a larger system size.
Figure 8
Comparison of the calculation results and predictions using the HDNNP model based on the SOAP descriptors for the aC dataset. The red (blue) dots represent the comparison results for the test (training) data. The difference in (a) and (b) are owing to differences in hyperparameters used for constructing SOAP descriptors. For (a), cutoff=6.0 Å, sigma=0.5, nmax=8, and lmax=8. For (b), sigma=1.0, while the other parameters are the same as those for (a).
Figure 1

Growing sequence of topological space

Data points

Filtration

Persistence diagram
Figure 2

a) 

b) 

Figure 3

a) 

b)
Figure 4

a) SOAP

b) ACSF

c) SOAP (average)

d) Coulomb matrix (CM)

e) CM (eigenspectrum)

f) Persistent Image
Figure 5

a) aGr

b) aC

c)  

d)  

\[ \begin{array}{c}
\text{input} \\
\text{PD (128x128)} \\
\text{conv} \\
\text{Max Pool} \\
\text{conv} \\
\text{Max Pool} \\
\text{conv} \\
\text{Max Pool} \\
\text{FC} \\
\text{FC} \\
\text{output} \\
\text{Energy} \\
\end{array} \]

\[ \begin{array}{c}
64 \\
(3x3) \\
64 \\
(3x3) \\
32 \\
(3x3) \\
8192 \\
4096 \\
\end{array} \]
Figure 6

(a) aGr

- RMSE (test) 8.3 meV/atom

(b) aC

- RMSE (test) 33.3 meV/atom

(c) 2,720 atoms

- RMSE 3.1 meV/atom

(d) 512 atoms

- RMSE 33.6 meV/atom
Figure 7

a)

Figure 8

a) Hyperparameters set1

b) Hyperparameters set2

RMSE (test) 108.6 meV/atom

RMSE (test) 29.2 meV/atom