PLOS ONE

RESEARCH ARTICLE

Estimation and prediction of ellipsoidal molecular shapes in organic crystals based on ellipsoid packing

Daiki Ito[1], Raku Shirasawa[2], Yoichiro Iino[2], Shigetaka Tomiya[2], Gouhei Tanaka[1]

1 Department of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan, **2** Materials Analysis Center, Fundamental Technology Research and Development Division 2, R&D Center, Sony Corporation, Atsugi, Japan

* gtanaka@g.ecc.u-tokyo.ac.jp

OPEN ACCESS

Citation: Ito D, Shirasawa R, Iino Y, Tomiya S, Tanaka G (2020) Estimation and prediction of ellipsoidal molecular shapes in organic crystals based on ellipsoid packing. PLoS ONE 15(9): e0239933. https://doi.org/10.1371/journal.pone.0239933

Editor: P. Davide Cozzoli, University of Salento, ITALY

Received: March 3, 2020

Accepted: September 16, 2020

Published: September 30, 2020

Data Availability Statement: All relevant data are within the manuscript and its supporting information file, except for the non-free material structure data used for our experiments. The identifiers of the material structure data in the Cambridge Structural Database (CSD) provided by Cambridge Crystallographic Data Centre (CCDC) are available upon a request to the R&D Center, Sony Corporation (contact address: Raku.Shirasawa@sony.com).

Abstract

Crystal structure prediction has been one of the fundamental and challenging problems in materials science. It is computationally exhaustive to identify molecular conformations and arrangements in organic molecular crystals due to complexity in intra- and inter-molecular interactions. From a geometrical viewpoint, specific types of organic crystal structures can be characterized by ellipsoid packing. In particular, we focus on aromatic systems which are important for organic semiconductor materials. In this study, we aim to estimate the ellipsoidal molecular shapes of such crystals and predict them from single molecular descriptors. First, we identify the molecular crystals with molecular centroid arrangements that correspond to affine transformations of four basic cubic lattices, through topological analysis of the dataset of crystalline polycyclic aromatic molecules. The novelty of our method is that the topological data analysis is applied to arrangements of molecular centroids intead of those of atoms. For each of the identified crystals, we estimate the intracrystalline molecular shape based on the ellipsoid packing assumption. Then, we show that the ellipsoidal shape can be predicted from single molecular descriptors using a machine learning method. The results suggest that topological characterization of molecular arrangements is useful for structure prediction of organic semiconductor materials.

Introduction

Finding novel materials with desired properties often requires exhaustive search. In computational materials science, *ab initio* calculations based on density functional theory (DFT) have played a central role in analyzing physical properties of materials and testing the validity of experimental results. Although *ab initio* calculations are powerful, versatile, and efficient, they are still computationally expensive for several important classes of problems [1]. An alternative approach is materials informatics which exploits data science and informatics for reducing computational cost in material research [2, 3]. In particular, machine learning techniques have been increasingly leveraged to identify the hidden rules governing the structure-property-

Funding: Funding for this study was provided by Sony Corporation. The funder provided support in the form of salaries for authors (RS, YI, and ST), but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: Funding for this study was provided by Sony Corporation. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

function relationship in materials from data. These methods have been successful in predicting material properties from atomistic and molecular information [4–12].

One of the challenging problems in materials science is crystal structure prediction (CSP). The goal of CSP is to accurately predict plausible crystal structures from atomistic and/or molecular information. The properties of molecular crystalline materials, such as energies and electronic characteristics, are highly sensitive to the arrangement of molecules due to complex intra- and inter-molecular interactions [13]. Therefore, CSP for molecular crystals is a significant step for materials property prediction [14]. Developing new effective computational methodologies for CSP is imperative for crystal engineering [15, 16], which aims to design crystalline materials with target structures leading to desired physical properties. However, even state-of-the-art computational methods for CSP require high computational cost for identifying plausible molecular arrangements which correspond to minimum energy structure [17].

From a geometrical viewpoint, some organic crystal structures are characterized by molecular packing [18, 19]. The structures of organic compounds with low-symmetry molecules have been analyzed under the close packing principle which implies that the minimum energies of compounds correspond to the packing structures of 3D molecular bodies with the least occupied volume [20–22]. The close packing principle and its modifications work well for many classes of organic crystals [23]. Whereas inorganic crystal structures composed of highly symmetric atomic bodies are often explained by close packing of spheres, organic crystal structures consisting of low-symmetry molecular bodies are linked to dense packing of ellipsoids [24, 25].

Many organic semiconductor materials are involved in aromatic systems. Motivated by this fact, we focus on the dataset of polycyclic aromatic molecules which are used to assess the effectiveness of our method. We first employ topological data analysis to identify the organic molecular crystals where the arrangement of molecular centroids coincides with affine transformations of basic cubic lattices. Then, we estimate the shapes of ellipsoids packed in the identified organic crystals under the ellipsoid packing assumption. Moreover, we show that the ellipsoidal shapes can be predicted from single molecular descriptors using a machine learning method for the dataset of polycyclic aromatic molecules. The ellipsoid radii can correspond to the approximate intracrystalline molecular shape, and therefore, our method is useful for predicting partial information of crystal structures.

Methods

Overview

An overview of our method for estimating ellipsoidal molecular shapes is shown in Fig 1. Our method relies on the concept of ellipsoid packing for organic crystal structures. We generate a persistence diagram from an arrangement of molecular centroids. If the generated persistence diagram coincides with a theoretically derived diagram for a basic cubic lattice, then we can estimate the ellipsoidal shapes from the identified lattice type and the specified affine transformation. The estimated ellipsoidal shape can be an approximate representation of an intracrystalline molecular shape, which is predicted from single molecular descriptors using a machine learning technique.

Ellipsoid packing

Crystal structures have been interpreted in terms of packing of atoms and molecules in crystallography. Some inorganic crystal structures have been explained as close packing of spheres

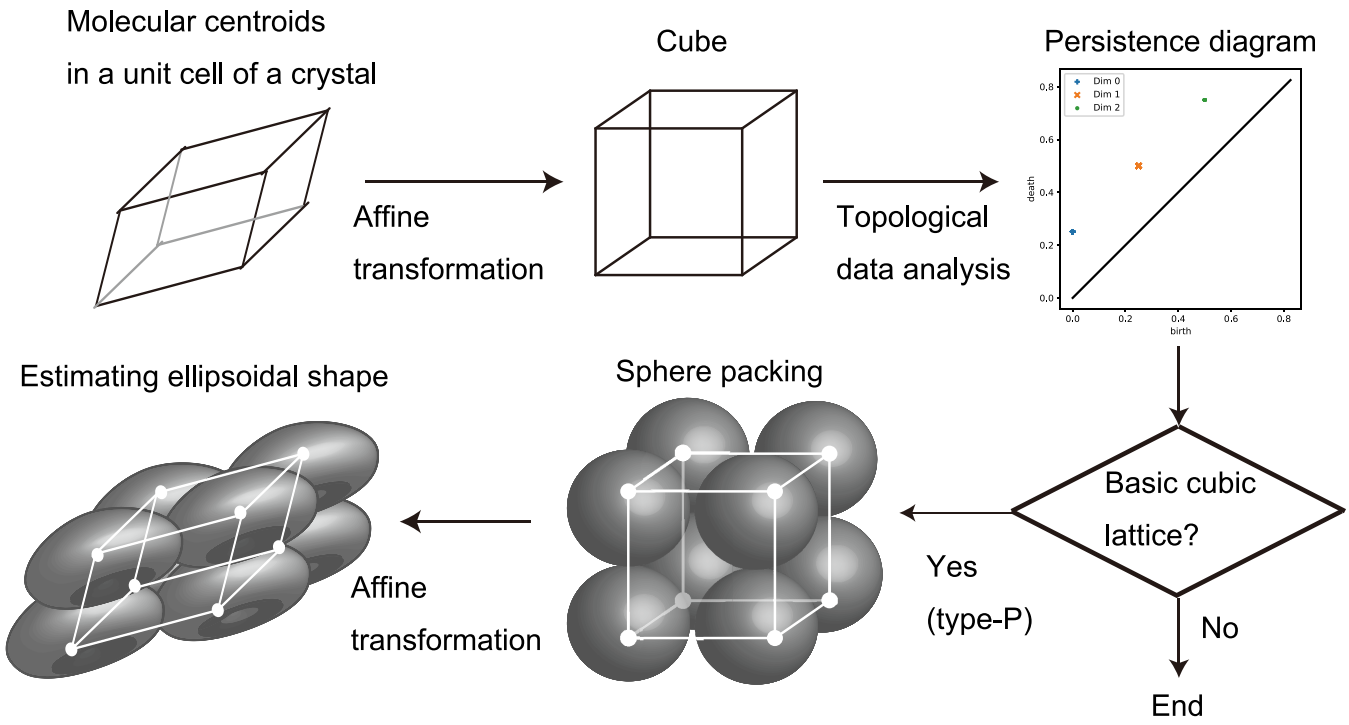

Fig 1. Methodology overview. An overview of the method to estimate the shapes of ellipsoids packed in molecular crystals is illustrated. This is an example for an organic molecular crystal where the molecular arrangement corresponds to the primitive (type-P) lattice (see Figs 2(a) and 4(a) for details).

https://doi.org/10.1371/journal.pone.0239933.g001

which approximate atomic bodies with high symmetry. The two models that achieve dense packing of identical spheres are known to be cubic closest packing (ccp) and hexagonal closest packing (hcp) [21, 22]. The structure of ccp is also known as a face-centered cubic (fcc) lattice in the cubic crystal system. For both ccp and hcp, the packing fraction is given by $\rho = \pi/\sqrt{18} \sim 0.74048$. Sphere packing itself has a long history in mathematics [26]. As for organic crystals, the constituent units are molecules with less symmetry and their bodies are suitably approximated by ellipsoids rather than spheres. Ellipsoid packing is an extended problem of sphere packing [27]. It was reported that densest packing structures of identical ellipsoids are limited to affine transformations of closest packing structures of identical spheres [24].

In this study, we deal with organic crystal structures from the viewpoint of ellipsoid packing following the above report, although an unusual case of densest crystal ellipsoid packing was later found in the glassy phases [28] and also in crystal packing. Molecular arrangements are determined by arrangements of centroids of molecules. We limit our focus to the crystals with molecular arrangements that are obtained by affine transformations of the cubic lattices illustrated in Fig 2. Fig 2(a)–2(d) correspond to the basic lattices, called Primitive (type-P), Base-centered (type-C), Body-centered (type-I), and Face-centered (type-F), respectively, in crystal systems. Affine transformations of these cubic lattices cover all the crystal families except for the hexagonal family. The hexagonal family is not considered in this study because the targeted crystal structure dataset does not contain crystal structures corresponding to its affine transformations. We first convert a unit cell of a molecular crystal to a cube by an affine transformation and then analyze it using persistent homology.

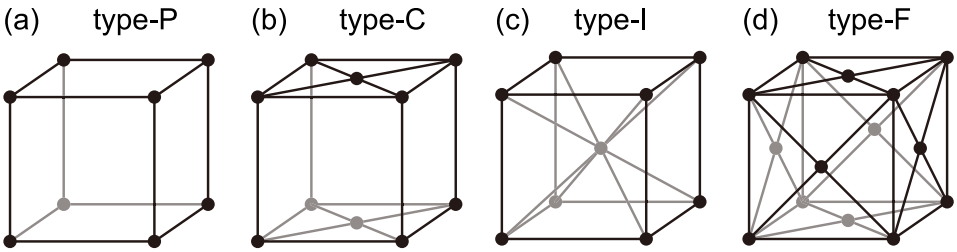

Fig 2. Basic cubic lattices. (a) Primitive (type-P) lattice. (b) Base-centered (type-C) lattice. (c) Body-centered (type-I) lattice. (d) Face-centered (type-F) lattice.

https://doi.org/10.1371/journal.pone.0239933.g002

Topological data analysis

Topological data analysis is an emerging mathematical technology to analyze topological properties of structural data using applied algebraic topology and computational geometry [29–31]. It enables to characterize qualitative features of a set of discrete points in space. Persistent homology is a powerful framework for topological data analysis [32, 33], which can reveal topological properties of a point cloud (i.e. a set of data points) at different spatial resolutions and generate a persistence diagram (i.e. a visualization of persistent homology as a 2D histogram).

Fig 3(a) illustrates an example of filtration for a point cloud on a 2D space. We consider disks with radius $r > 0$, centered at the four data points. As r is increased from 0, the initially separated disks start to overlap with each other at a certain value of r. The change in the r value means a change in the resolution. If the union of disks makes a hole at $r = b$ and the hole vanishes at $r = d$ as in Fig 3(a), then the birth and death of the hole are recorded as a point at (b, d) in the persistence diagram as shown in Fig 3(b). Such points are plotted with respect to each hole that appears when continuously increasing the disk size. Plotted points in a persistence

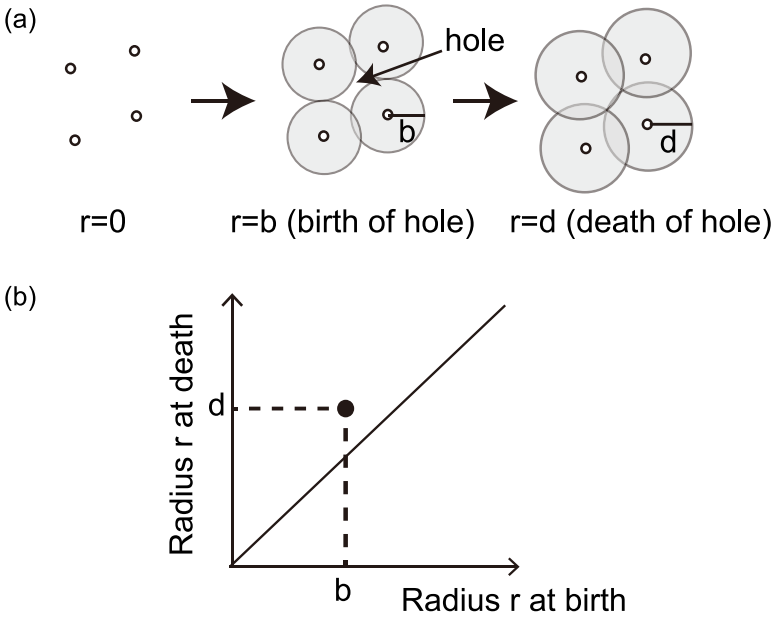

Fig 3. An example of persistent homology. (a) Filtration for a point cloud on a 2D space. (b) Persistence diagram.

https://doi.org/10.1371/journal.pone.0239933.g003

diagram exist above the diagonal line, because a death of a hole occurs after its birth. Topologically similar point clouds produce similar persistence diagrams.

The persistence diagram for a point cloud $P = \{\mathbf{p}_i \in \mathbb{R}^3 \mid i = 1, \ldots, I\}$ consisting of I discrete points in a 3D space is defined as follows:

$$D_q(P) \quad = \quad \{(b_j, d_j) \in \mathbb{R}^2 \mid j = 1, \ldots, J\}, \tag{1}$$

where J is the number of holes and q is the dimensionality of holes, e.g. $q = 0$ for connected components, $q = 1$ for rings, and $q = 2$ for cavities. We use the multi-set $\{D_q(P) \mid q = 0, 1, 2\}$ to characterize a topological feature of the point cloud P.

We analytically derived the persistence diagrams for the four cubic lattices shown in Fig 2, where the length of each side is given by α. Denoting the point clouds for type-P, type-C, type-I, and type-F lattices by P_{P}, P_{C}, P_{I}, and P_{F}, respectively, we obtain the following results:

- For type-P,

$$\begin{aligned} D_0(P_{\mathrm{P}}) &= \{(0, \alpha/2), (0, \infty)\}, \\ D_1(P_{\mathrm{P}}) &= \{(\alpha/2, \alpha/\sqrt{2})\}, \\ D_2(P_{\mathrm{P}}) &= \{(\alpha/\sqrt{2}, \sqrt{3}\alpha/2)\}. \end{aligned} \tag{2}$$

- For type-C,

$$\begin{aligned} D_0(P_{\mathrm{C}}) &= \{(0, \alpha/2\sqrt{2}), (0, \alpha/2), (0, \infty)\}, \\ D_1(P_{\mathrm{C}}) &= \{(\alpha/2\sqrt{2}, \alpha/2), (\alpha/2, \sqrt{3}\alpha/2\sqrt{2})\}, \\ D_2(P_{\mathrm{C}}) &= \{(\sqrt{3}\alpha/2\sqrt{2}, \alpha/\sqrt{2})\}. \end{aligned} \tag{3}$$

- For type-I,

$$\begin{aligned} D_0(P_{\mathrm{I}}) &= \{(0, \sqrt{3}\alpha/4), (0, \infty)\}, \\ D_1(P_{\mathrm{I}}) &= \{(\sqrt{3}\alpha/4, 3\alpha/4\sqrt{2}), (\alpha/2, 3\alpha/4\sqrt{2})\}, \\ D_2(P_{\mathrm{I}}) &= \{(3\alpha/4\sqrt{2}, \sqrt{5}\alpha/4)\}. \end{aligned} \tag{4}$$

- For type-F,

$$\begin{aligned} D_0(P_{\mathrm{F}}) &= \{(0, \alpha/2\sqrt{2}), (0, \infty)\}, \\ D_1(P_{\mathrm{F}}) &= \{(\alpha/2\sqrt{2}, \alpha/\sqrt{6})\}, \\ D_2(P_{\mathrm{F}}) &= \{(\alpha/\sqrt{6}, \sqrt{3}\alpha/4), (\alpha/\sqrt{6}, \alpha/2)\}. \end{aligned} \tag{5}$$

The corresponding persistence diagrams are shown in Fig 4(a)–4(d).

The similarity between two persistence diagrams, X and Y, can be measured with the following bottleneck distance [29]:

$$d_B(X, Y) \quad = \quad \inf_{\eta: X \to Y} \sup_{x \in X} ||x - \eta(x)||_\infty, \tag{6}$$

where η is a bijection between X and Y, and the L_∞-distance between points $u = (u_1, u_2)$ and

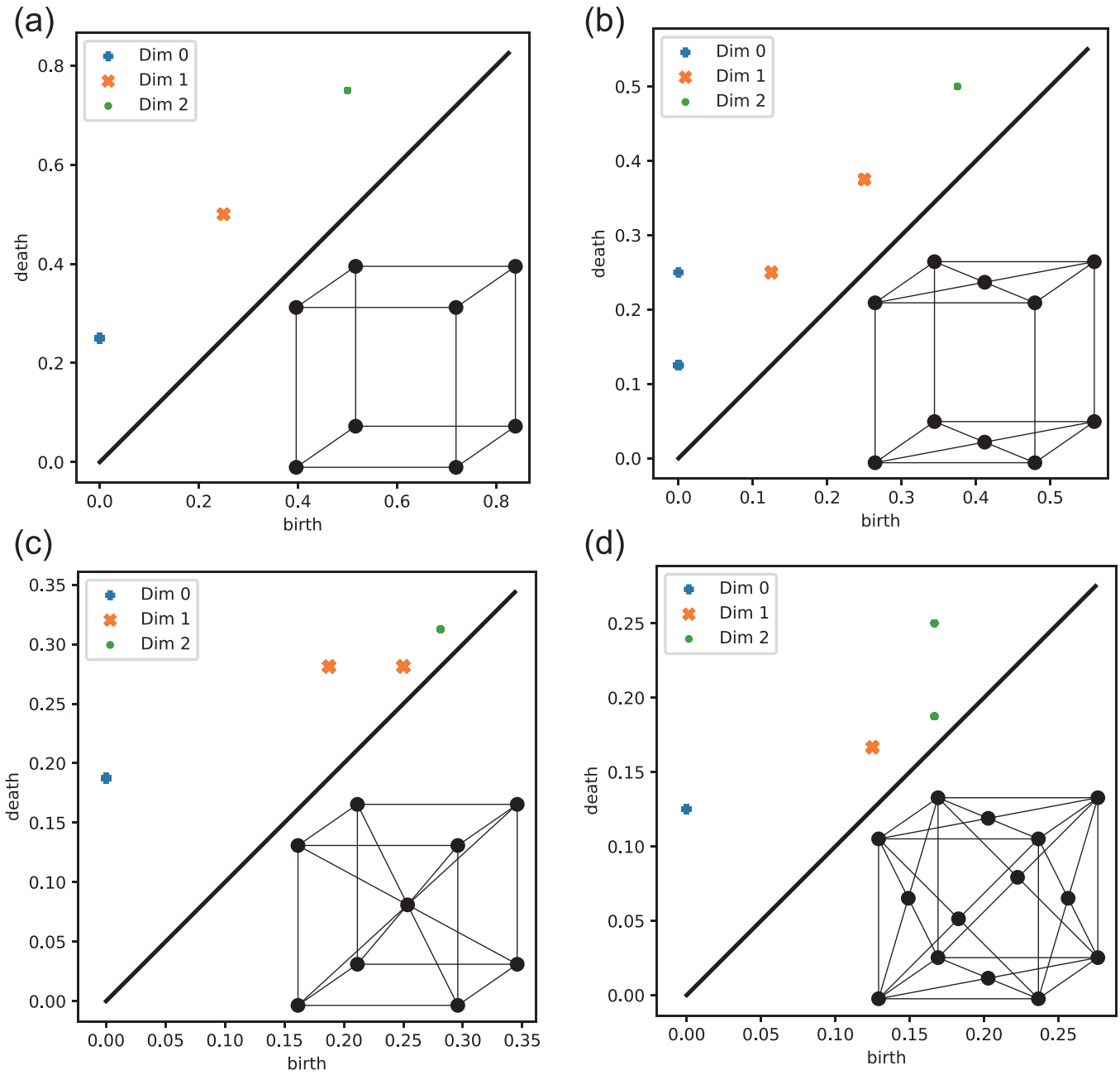

Fig 4. Persistent diagrams for the cubic lattices in Fig 2. In each panel, the blue pluses, orange crosses, and green dots indicate points for $q = 0, 1, 2$. In the right-bottom part of each panel, the corresponding lattice structures are shown. (a) type-P (Eq (2)). (b) type-C (Eq (3)). (c) type-I (Eq (4)). (d) type-F (Eq (5)).

https://doi.org/10.1371/journal.pone.0239933.g004

$v = (v_1, v_2)$ is defined as follows:

$$||u - v||_\infty = \max\{|u_1 - v_1|, |u_2 - v_2|\}. \quad (7)$$

In short, the bottleneck distance is the cost of the optimal matching between points of the two diagrams.

Using this similarity measure, we identify the crystals where the molecular arrangements correspond to affine transformations of the cubic lattices in Fig 2. First, a unit cell of a crystal,

represented as a parallelepiped, is converted to a cube with each side $\alpha = 0.1$ nm by an affine transformation. This normalization transformation is necessary because a persistence diagram is robust against a rotation of the targeted point cloud and noise but largely affected by a difference in the scales. Then, we compute persistence diagrams $D_q(P_M)$ ($q = 0, 1, 2$) for a point cloud P_M in the cube, corresponding to a set of molecular centroids. If the multi-set $D_q(P_M)$ is sufficiently close to one of those in Eqs (2)–(5), then the arrangement of molecular centroids is categorized into the same affine group. The persistence diagram for P_M is considered to be equivalent to that of P_M (Z = P, C, I, or F), if the following inequality is satisfied:

$$\min_{q=0,1,2} d_B(D_q(P_M), D_q(P_Z)) < \epsilon, \tag{8}$$

where ϵ represents the acceptable error. In later numerical experiments, we set ϵ at 0.001 nm. For numerically generating persistence diagrams for molecular centroids, we used Dionysus 2 which is a library for computing persistent homology [34].

Estimation method of ellipsoidal shapes

Once we identify a crystal that has a structure corresponding to an affine transformation of a cubic lattice, we can estimate the shape of identical ellipsoids around the molecular centroids packed in the identified cubic lattice. In general, an ellipsoid is expressed as follows:

$$\mathbf{x}^\top R\mathbf{x} = 1, \tag{9}$$

where $\mathbf{x} = (x_1, x_2, x_3)^\top$ is a 3D coordinate. The diagonal matrix R is represented as follows:

$$R = \begin{bmatrix} r_1^{-2} & 0 & 0 \\ 0 & r_2^{-2} & 0 \\ 0 & 0 & r_3^{-2} \end{bmatrix}, \tag{10}$$

where r_i denotes the radius in the direction of x_i for $i = 1, 2, 3$.

For ellipsoids (or spheres) packed in the four cubic lattices, the diagonal matrices R_Z for Z = P, C, I, or F are calculated as follows:

$$R_P = \begin{bmatrix} (\alpha/2)^{-2} & 0 & 0 \\ 0 & (\alpha/2)^{-2} & 0 \\ 0 & 0 & (\alpha/2)^{-2} \end{bmatrix}, \tag{11}$$

$$R_C = \begin{bmatrix} (\sqrt{2}\alpha/4)^{-2} & 0 & 0 \\ 0 & (\sqrt{2}\alpha/4)^{-2} & 0 \\ 0 & 0 & (\alpha/2)^{-2} \end{bmatrix}, \tag{12}$$

$$R_I = \begin{bmatrix} (\sqrt{3}\alpha/4)^{-2} & 0 & 0 \\ 0 & (\sqrt{3}\alpha/4)^{-2} & 0 \\ 0 & 0 & (\sqrt{3}\alpha/4)^{-2} \end{bmatrix}, \tag{13}$$

$$R_{\mathrm{F}} = \begin{bmatrix} (\sqrt{2}\alpha/4)^{-2} & 0 & 0 \\ 0 & (\sqrt{2}\alpha/4)^{-2} & 0 \\ 0 & 0 & (\sqrt{2}\alpha/4)^{-2} \end{bmatrix}, \quad (14)$$

where α is the length of each side of the unit cell.

Now we assume that an ellipsoid approximating a molecular shape in a crystal is represented as follows:

$$\mathbf{v}^{\top} Q \mathbf{v} = 1 \quad (15)$$

where $\mathbf{v}$ is a 3D coordinate and Q is an unknown diagonal matrix determining the shape of the ellipsoid. We denote by W the affine transformation used to normalize the unit cell of a crystal when generating the persistence diagram. Then the coordinate transformation is represented as $\mathbf{x} = W\mathbf{v}$. The ellipsoid in Eq (15) is transformed by W into the ellipsoid $\mathbf{x}^{\top} R_{\mathrm{Z}} \mathbf{x} = 1$ with a known diagonal matrix R_{Z} (Z = P, C, I, or F). By substituting $\mathbf{x} = W\mathbf{v}$ into $\mathbf{x}^{\top} R_{\mathrm{Z}} \mathbf{x} = 1$, we obtain

$$\mathbf{v}^{\top}(W^{\top} R_{\mathrm{Z}} W)\mathbf{v} = 1, \quad (16)$$

where $W^{\top} R_{\mathrm{Z}} W$ is a symmetric matrix. The radii r_i (i = 1, 2, 3) of the ellipsoid approximating the intracrystalline molecular shape are computed as $r_i = 1/\sqrt{\lambda_i}$, where λ_i (i = 1, 2, 3) represent the eigenvalues of $W^{\top} R_{\mathrm{Z}} W$.

Prediction method of ellipsoidal shapes

The above-mentioned method gives approximate ellipsoidal molecular shapes represented as ellipsoid radii. We test whether these ellipsoidal shapes can be predicted from single molecular descriptors using a machine learning method. Molecular descriptors represent a set of features of single molecules, which have many possible representations. Molecular fingerprint is one of the widely used descriptors to determine the similarity of chemical structures [35, 36]. A molecular fingerprint is represented as a binary feature vector where each component expresses whether an attribute is present or absent in the molecule.

We employ Extended Connectivity Fingerprint (ECFP) [37] which is one of the data-driven circular fingerprints unlike those based on predefined substructural keys. A molecule is represented as a graph where the vertices are atoms and the edges are bonds. Subgraphs included in the neighborhood of each vertex up to a fixed diameter are examined and quantified with the atomic features using the Morgan algorithm [38]. Then these substructural features are mapped into integer codes using a hashing procedure to keep the length of the feature vector fixed. The ECFP is obtained as a binary feature vector from the resulting identifiers. For converting molecular information into ECFPs, we used Chainer Chemistry which is an open-source library for deep learning in biology and chemistry [39].

The ellipsoidal shape prediction is performed with a feedforward neural network with one hidden layer in a supervised learning framework. The hidden layer has 64 units. Each node has the hyperbolic tangent (tanh) activation function. The number of training data is denoted by N. The nth teacher data is given by a pair of input and output, where the input is an ECFP for the molecule and the output is the estimated ellipsoid radii $r_1^{(n)}$, $r_2^{(n)}$, and $r_3^{(n)}$ for the corresponding molecular crystal. We train a neural network model so as to minimize the mean squared

error between the network output $\hat{r}_i^{(n)}$ and the teacher output $r_i^{(n)}$, described as follows:

$$E = \frac{1}{2N}\sum_{n=1}^{N}\sum_{i=1}^{3}\left(\hat{r}_i^{(n)} - r_i^{(n)}\right)^2. \quad (17)$$

Using a test dataset, we evaluate the prediction accuracy in predicting the ellipsoidal shape.

Results

Dataset

Molecular crystals are versatile materials which can be found in pharmaceuticals, organic semiconductors, solid-state reactions, and plastic materials [40]. For instance, polycyclic aromatic hydrocarbons (PAHs) and their derivatives have been widely explored for organic semiconductors [41]. Such organic molecular crystals are important targets of CSP because their molecular arrangements are deeply involved in electron mobility in the crystals [42]. We selected polyaromatic crystals that consist of only one type of molecule from the Cambridge Structural Database (CSD) provided by Cambridge Crystallographic Data Centre (CCDC) [43]. We made the following three crystal datasets:

- Polycyclic aromatic hydrocarbons (PAHs): 75 crystals of polycyclic aromatic hydrocarbons that contain only hydrogens and carbons.
- Polycyclic aromatics with hetero atoms (PAHAs): 404 crystals of polycyclic aromatic hydrocarbons that can contain hetero elements (N, S, etc.) in their skeletons but have no substituents other than halogens.
- Polycyclic aromatics (PAs): 8787 crystals of polycyclic aromatic hydrocarbons that can contain hetero elements in their skeletons and arbitrary substituents.

PAHs are included in PAHAs which are included in PAs, as shown in Fig 5.

Identification of crystals with specific structures

The topological data analysis was applied to molecular centroids of each crystal structure in the datasets. It was determined whether each crystal structure corresponds to an affine transformation of one of the four cubic lattices shown in Fig 2. Table 1 shows the numbers of organic crystals that were classified into the four types (i.e. type-P, type-C, type-I, and type-F) for each dataset. The others were categorized into the class of "Others." The fraction of crystals corresponding to the four types to the total number of crystals in the PAHs is about 30%, that in the PAHAs is about 21%, and that in the PAs is about 7%. The high fraction of simple crystal structures in the PAHs is related to the fact that some crystals in the PAHs tend to have layered molecular arrangements and their structures are classified into typical packing motifs [44, 45]. It also implies that the crystal structure is complex when hetero atoms are contained in molecules as in the PAHAs and PAs.

Fig 6 illustrates the examples of molecular centroid arrangements and the corresponding persistence diagrams for the crystals categorized in the five types (see Table 1). Fig 6(a)–6(d) show 9,10-bis(2-(4-(n-Decyloxy)phenyl)vinyl)anthracene with type-P structure, Pentacene with type-C structure, Anthracene with type-I structure, and Benzene with type-F structure, respectively. In other words, these persistence diagrams are found in Fig 4. Fig 6(e) shows Acenaphthobenzopicene which has a lower symmetric structure than the four other crystal structure types as seen from the dispersion of points in the persistence diagram.

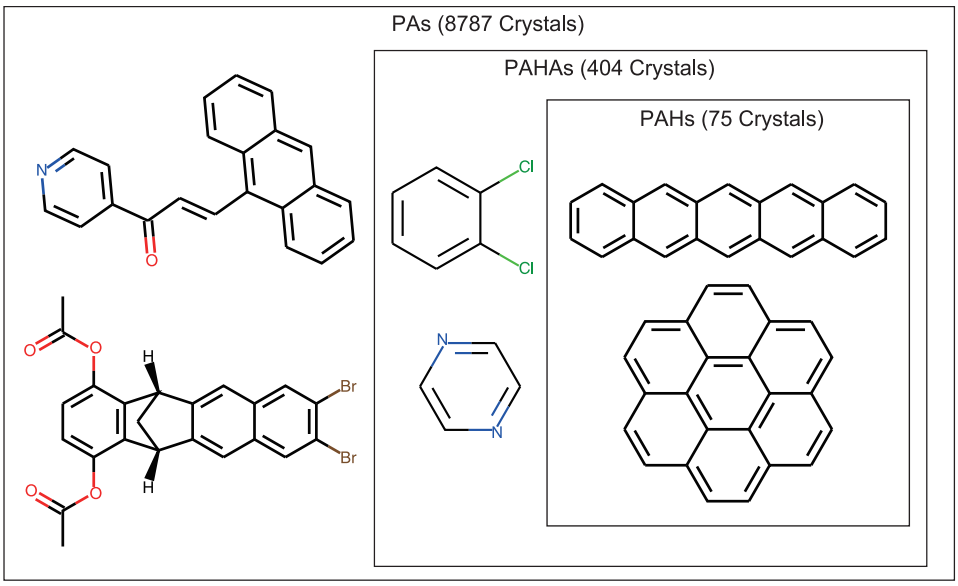

Fig 5. Dataset of organic molecular crystals. The three datasets correspond to polycyclic aromatic hydrocarbons (PAHs), polycyclic aromatics with hetero atoms (PAHAs), and polycyclic aromatics (PAs).

https://doi.org/10.1371/journal.pone.0239933.g005

Ellipsoidal shape estimation

We assume that identical ellipsoids around the molecular centroids are packed in the crystals that were identified in Table 1. The ellipsoidal shape was estimated from Eq (16). Fig 7 shows the distribution of the estimated ellipsoidal radii for the PAs that correspond to the affine transformations of the basic cubic lattices. The three axes indicate r_1, r_2, and r_3, satisfying $r_1 \geq r_2 \geq r_3$. The different marks correspond to different types of cubic lattices in Fig 2. To validate the above assumption, we investigated a correlation between the ellipsoidal volume $V = (4/3)\pi r_1 r_2 r_3$ calculated from the estimated radii and the molecular volume calculated based on the electron density through Monte-Carlo integration in Gaussian 16 [46]. The results are shown in Fig 8. The molecular structures and the electron densities were calculated by DFT calculation at the theoretical level of B3LYP with 6-31G basis set. The result shows a high correlation (the Pearson's correlation coefficient: 0.897), indicating that the ellipsoidal volume well approximates the molecular volume. Similarly, we obtained the estimated ellipsoid radii for the two other datasets, PAHAs and PAHs (not shown).

The estimated ellipsoid radii for some crystals are listed in Table 2. The structures of the molecules listed in Table 2(a)–2(g) are shown in Fig 9(a)–9(g). It shows that the ellipsoidal shapes are affected by the single molecular structures. For example, Benzene (Fig 9(a)), Anthracene (Fig 9(b)), and Pentacene (Fig 9(c)) have one, three, and five aromatic rings in a

Table 1. Identification of crystals corresponding to affine transformed cubic lattices.

	PAHs	PAHAs	PAs
type-P	0	0	160
type-C	9	34	205
type-I	8	34	153
type-F	6	15	104
Others	52	321	8165
Total	75	404	8787

https://doi.org/10.1371/journal.pone.0239933.t001

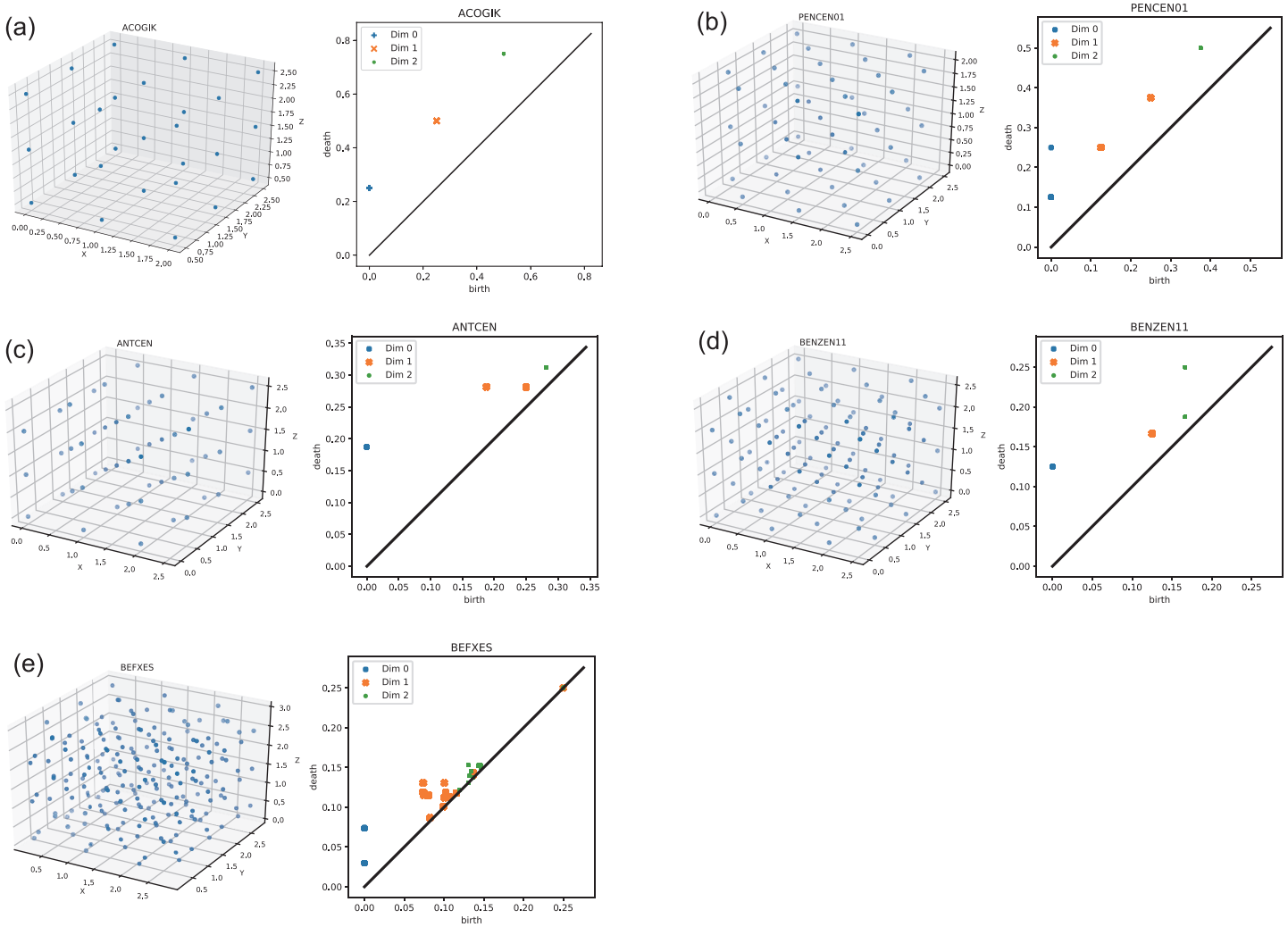

Fig 6. Examples of persistence diagrams for organic crystals. Molecular centroids (left) and corresponding persistence diagrams (right). In each persistence diagram, the blue pluses, orange crosses, and green circles indicate the plots of D_q with $q = 0$, $q = 1$, and $q = 2$, respectively. (a) 9,10-bis(2-(4-(n-Decyloxy)phenyl)vinyl)anthracene (PAs) with type-P structure. (b) Pentacene (PAHs) with type-C structure. (c) Anthracene (PAHs) with type-I structure. (d) Benzene (PAHs) with type-F structure. (e) Acenaphthobenzopicene (PAHs) which is categorized into "Others".

https://doi.org/10.1371/journal.pone.0239933.g006

series, respectively. As the number of aromatic rings increases, only the estimated radius r_1 tends to be elongated. Benzophenanthrene (Fig 9(d)) and Tetrabenzocoronene (Fig 9(e)) have planar disk-like structures, and thus, the estimated values of r_1 and r_2 are relatively large compared with that of r_3. While the single molecular structure is a major factor affecting the ellipsoid radii, its chemical structure also influences them. The shapes of Pyrazine (Fig 9(f)) and 2,7-bisacridine (Fig 9(g)) look similar to the shapes of Benzen and Anthracene, respectively, but r_1 is longer due to the influence of terminal hetero atoms.

Ellipsoidal shape prediction

We performed a machine learning prediction of the ellipsoidal radii for PAs from single molecular information. This is regarded as a simplified task of CSP, because the information of arrangements of molecular centroids and molecular shapes are useful for identifying crystal structures. The task is to predict the radii of ellipsoids (r_1, r_2, and r_3) from molecular

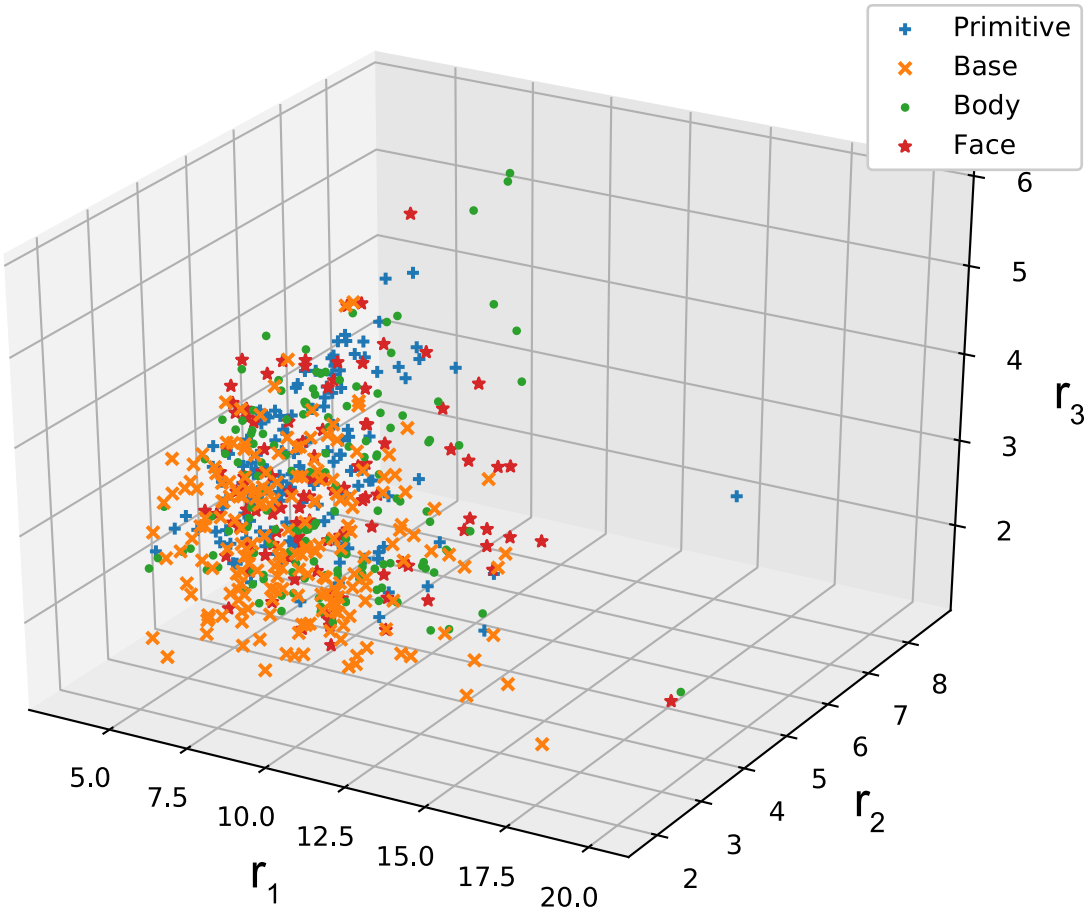

Fig 7. Estimated radii of ellipsoids. Each plot at (r_1, r_2, and r_3) indicates the radii of ellipsoids packed in a crystal that was identified in Table 1. The blue pluses, orange crosses, green circles, and red stars correspond to type-P, type-C, type-I, and type-R, respectively.

https://doi.org/10.1371/journal.pone.0239933.g007

fingerprints given by ECFPs. We trained a neural network model by using the Adam (Adaptive moment estimation) optimizer with learning rate 0.001 and mini-batch size 32 [47]. We used 10% of the training data for early stopping to avoid overfitting and evaluated our model via four-fold cross validation. The results of the ellipsoid radii prediction are shown in Fig 10. In each panel, The horizontal axis represents the true radius and the vertical axis represents the predicted one. The plots for the r_1 prediction are close to the diagonal line as shown in Fig 10(a), implying a successful prediction. The mean training error is around 0.04 nm and the testing error is around 0.094 nm. On the other hand, we can find that the predicted values for r_2 and r_3 tend to be smaller than the actual values when their values are large as shown in Fig 10(b) and 10(c). The prediction results for PAHs and PAHAs are shown in S1 Fig. The results suggest that our method is useful for predicting the length of the main axis of the skeleton in organic semiconductor materials rather than lengths of the shorter ellipsoid half-axes. As seen in Table 2, the ellipsoid radii are influenced not only by the single molecular shape but also by the chemical constitution. Therefore, the molecular fingerprints including such information worked well in the ellipsoidal shape prediction. We can choose other machine learning models for the radii prediction and an improvement in the prediction performance is an issue to be considered.

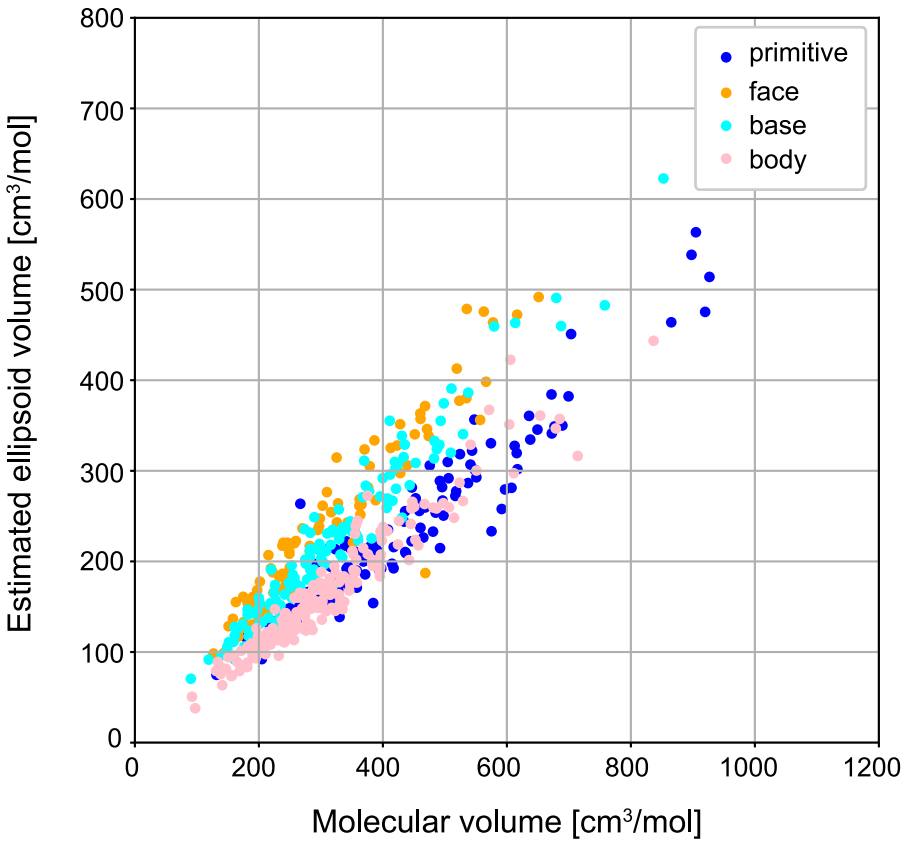

Fig 8. Correlation between the estimated ellipsoid volume and the molecular volume. The ellipsoid volume [cm^3/mol] (the vertical axis) was calculated from the estimated radii. The molecular volume [cm^3/mol] (the horizontal axis) was calculated using Gaussian 16 which is a general purpose computational chemistry software [46]. The Pearson's correlation coefficient for all the points is 0.897.

https://doi.org/10.1371/journal.pone.0239933.g008

Conclusion and discussion

We have studied the problem of ellipsoid estimation and prediction in terms of molecular packing, mainly targeted for polycyclic aromatics found in organic semiconductor materials. We have identified the organic molecular crystals whose structures are characterized by affine transformations of the four basic cubic lattices, through the topological analysis of the dataset of molecular centroids in crystals of aromatic molecules. Then, we have computed the radii of ellipsoids around the molecular centroids of those crystals from the identified lattice type and the specified affine transformation under the dense packing assumption. The ellipsoid shape

Table 2. Examples of estimated ellipsoid radii.

Molecules	r_1	r_2	r_3
(a) Benzene	0.3253 nm	0.2576 nm	0.2365 nm
(b) Anthracene	0.5469 nm	0.2699 nm	0.2699 nm
(c) Pentacene	0.7142 nm	0.2707 nm	0.2190 nm
(d) Benzophenanthrene	0.5183 nm	0.5005 nm	0.2045 nm
(e) Tetrabenzocoronene	0.8702 nm	0.6280 nm	0.1630 nm
(f) Pyrazine	0.4041 nm	0.2532 nm	0.1634 nm
(g) 2,7-bisacridine	0.9566 nm	0.2483 nm	0.2144 nm

https://doi.org/10.1371/journal.pone.0239933.t002

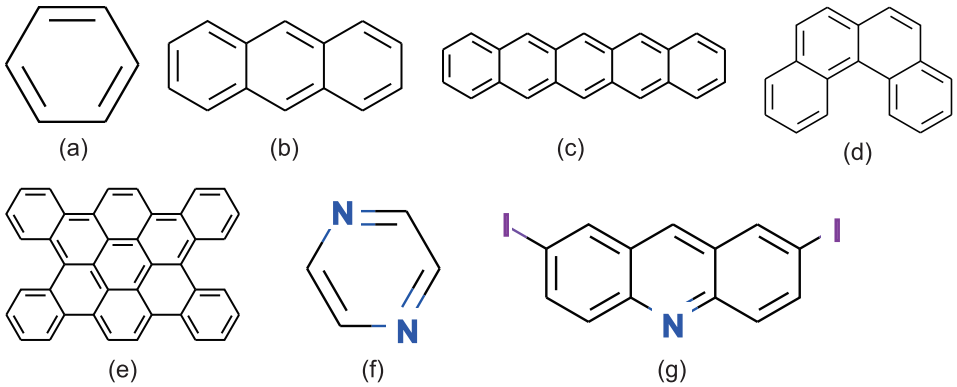

Fig 9. Structures of single isolated molecules in Table 2. (a) Benzene. (b) Anthracene. (c) Pentacene. (d) Benzophenanthrene. (e) Tetrabenzocoronene. (f) Pyrazine. (g) 2,7-bisacridine.

https://doi.org/10.1371/journal.pone.0239933.g009

represents an approximate shape of the intracrystalline molecules, which provides partial information of molecular structures in crystals. In additional experiments, we have shown that the ellipsoid radii can be predicted from the single molecular descriptors. Our results suggest that a combination of topological data analysis and machine learning can partly contribute to CSP.

It has been known that some molecular crystal structures are explained by ellipsoid packing corresponding to minimum energy structure. However, it was not straightforward to automatically identify such crystals from molecular centroid data because of the difficulty in checking coincidence of point groups in a 3D space. Thus, we have used the persistent homology to transform the point arrangement in the 3D space into the 2D persistence diagram. This method is useful particularly when checking the topological similarity between molecular arrangements. Our experiments have shown that some molecular arrangements (centroids) can be classified into basic lattices via affine transformations. This is analogous to the fact that spatial arrangements of atoms in crystals are characterized by Bravais lattices. A further study on topological classification of molecular arrangements would be effective for understanding molecular structures in crystals.

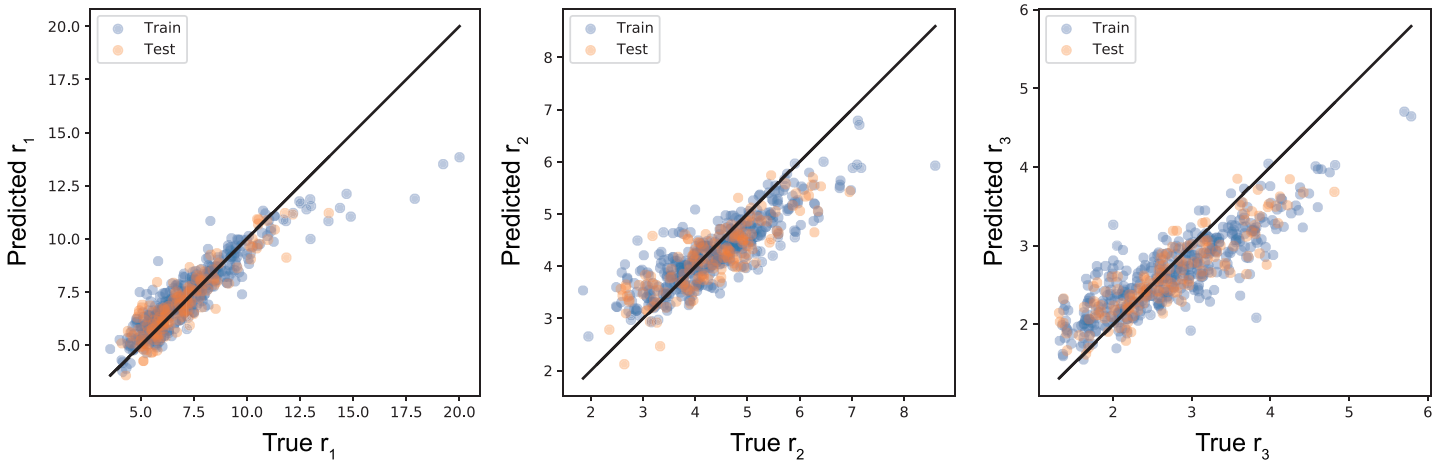

Fig 10. Prediction of the ellipsoid radii for PAs from ECFPs. The horizontal axis represents the teacher data and the vertical axis represents the predicted value. (a) r_1. (b) r_2. (c) r_3.

https://doi.org/10.1371/journal.pone.0239933.g010

The focus of our study has been limited to the crystals with specific structures obtained by affine transformations of the basic cubic lattices, which cover triclinic, monoclinic, orthorhombic, tetragonal, and cubic crystal families. The remaining family is the hexagonal crystal family. We have excluded this family in this study due to the difficulty in analytical derivation of the corresponding persistence diagram and the ellipsoid radii under the dense packing condition. A future work is to extend our method such that the crystals having structures corresponding to affine transformations of hexagonal lattices are also handled. Another issue is that the affine transformation restricts the molecular orientations. To distinguish diverse orientations such as a herringbone motif, it would be necessary to apply other transformations between molecular centroids and basic lattices.

The proposed method has several limitations. First, it does not give information about the arrangements of atoms in each crystalline molecule. For a full CSP, a conformational search of the arrangements of atoms after applying the proposed method would be necessary. Second, the assumptions in our method are applicable only to a part of crystal structures. For instance, the approximation of molecules with ellipsoids would be valid for rigid molecules but not for highly flexible molecules. A possible approach for examining how much molecular flexibility is permitted is to investigate the effect of the number of dihedral angles, representing the flexibility level, on the prediction accuracy. The dense packing assumption would also not be applicable to structures with substituents such as -CHO hydrogen bond, because they can be low-density structures. For such structures, *ab initio* calculations with force fields and DFT would be appropriate [48]. Third, the fraction of the number of structures identified based on our assumption is not large as shown in Table 1 for the dataset used in this study. This might be caused by the smallness of the acceptable error ϵ in Eq (8), which is a severe condition for determining identical persistence diagrams. There is a possibility that the fraction is increased by increasing the value of the acceptable error ϵ in Eq (8), but there would still be remaining structures that are not identified by our assumption. To extend the applicability of our method, we need to consider other lattice types in addition to those in Fig 2.

Supporting information

S1 Fig. Prediction of the ellipsoid radii for PAHs and PAHAs from ECFPs. The horizontal axis represents the teacher data and the vertical axis represents the predicted value. (a)-(c) the prediction results on r_1, r_2, and r_3 for PAHs. (d)-(f) the prediction results on r_1, r_2, and r_3 for PAHAs.
(EPS)

Author Contributions

Conceptualization: Daiki Ito, Raku Shirasawa, Gouhei Tanaka.

Data curation: Daiki Ito, Raku Shirasawa.

Formal analysis: Daiki Ito.

Funding acquisition: Raku Shirasawa.

Investigation: Daiki Ito, Raku Shirasawa.

Methodology: Daiki Ito.

Project administration: Raku Shirasawa, Gouhei Tanaka.

Supervision: Shigetaka Tomiya, Gouhei Tanaka.

Visualization: Daiki Ito, Gouhei Tanaka.

Writing – original draft: Gouhei Tanaka.

Writing – review & editing: Raku Shirasawa, Yoichiro Iino, Shigetaka Tomiya.

References

1. Neugebauer J, Hickel T. Density functional theory in materials science. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2013; 3(5):438–448. https://doi.org/10.1002/wcms.1125 PMID: 24563665
2. Rajan K. Materials informatics. Materials Today. 2005; 8(10):38–45. https://doi.org/10.1016/S1369-7021(05)71123-8
3. Ward L, Wolverton C. Atomistic calculations and materials informatics: A review. Current Opinion in Solid State and Materials Science. 2017; 21(3):167–176. https://doi.org/10.1016/j.cossms.2016.07.002
4. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters. 2012; 108(5):058301. https://doi.org/10.1103/PhysRevLett.108.058301 PMID: 22400967
5. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. Scientific Reports. 2013; 3:2810. https://doi.org/10.1038/srep02810 PMID: 24077117
6. Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A, et al. Machine learning of molecular electronic properties in chemical compound space. New Journal of Physics. 2013; 15 (9):095003. https://doi.org/10.1088/1367-2630/15/9/095003
7. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Big data meets quantum chemistry approximations: the Δ-machine learning approach. Journal of Chemical Theory and Computation. 2015; 11 (5):2087–2096. https://doi.org/10.1021/acs.jctc.5b00099 PMID: 26574412
8. Ryan K, Lengyel J, Shatruk M. Crystal structure prediction via deep learning. Journal of the American Chemical Society. 2018; 140(32):10158–10168. https://doi.org/10.1021/jacs.8b03913 PMID: 29874459
9. Yamashita T, Sato N, Kino H, Miyake T, Tsuda K, Oguchi T. Crystal structure prediction accelerated by Bayesian optimization. Physical Review Materials. 2018; 2(1):013803. https://doi.org/10.1103/PhysRevMaterials.2.013803
10. Yang J, De S, Campbell JE, Li S, Ceriotti M, Day GM. Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction. Chemistry of Materials. 2018; 30 (13):4361–4371. https://doi.org/10.1021/acs.chemmater.8b01621
11. Honrao S, Anthonio BE, Ramanathan R, Gabriel JJ, Hennig RG. Machine learning of ab-initio energy landscapes for crystal structure predictions. Computational Materials Science. 2019; 158:414–419. https://doi.org/10.1016/j.commatsci.2018.08.041
12. Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y. Generative adversarial networks for crystal structure prediction. arXiv preprint arXiv:200401396. 2020.
13. Dunitz JD, Gavezzotti A. How molecules stick together in organic crystals: weak intermolecular interactions. Chemical Society Reviews. 2009; 38(9):2622–2633. https://doi.org/10.1039/b822963p PMID: 19690742
14. Day GM, Gorbitz CH. Introduction to the special issue on crystal structure prediction. Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials. 2016; 72:435–436. https://doi.org/10.1107/S2052520616012348 PMID: 27484366
15. Tiekink ER, Vittal J, Zaworotko M. Organic crystal engineering: frontiers in crystal engineering. John Wiley & Sons; 2010.
16. Campbell JE, Yang J, Day GM. Predicted energy–structure–function maps for the evaluation of small molecule organic semiconductors. Journal of Materials Chemistry C. 2017; 5(30):7574–7584. https://doi.org/10.1039/C7TC02553J
17. Reilly AM, Cooper RI, Adjiman CS, Bhattacharya S, Boese AD, Brandenburg JG, et al. Report on the sixth blind test of organic crystal structure prediction methods. Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials. 2016; 72(4):439–459. https://doi.org/10.1107/S2052520616007447 PMID: 27484368
18. Craig DP, Mason R, Pauling P, Santry D. Molecular packing in crystals of the aromatic hydrocarbons. Proceedings of the Royal Society of London Series A Mathematical and Physical Sciences. 1965; 286 (1404):98–116.

19. Williams DE, Starr TL. Calculation of the crystal structures of hydrocarbons by molecular packing analysis. Computers & Chemistry. 1977; 1(3):173–177. https://doi.org/10.1016/0097-8485(77)85007-9 PMID: 8003270

20. Kitaigorodskii AI. Organic chemical crystallography. Consultants Bureau, New York; 1961.

21. Kitaigorodskii A. The principle of close packing and the condition of thermodynamic stability of organic crystals. Acta Crystallographica. 1965; 18(4):585–590. https://doi.org/10.1107/S0365110X65001391

22. Kitaigorodsky A. Molecular Crystals and Molecules. Academic Press, New York; 1973.

23. Slovokhotov YL. Organic crystallography: three decades after Kitaigorodskii. Structural Chemistry. 2018; p. 1–8.

24. Matsumoto T, Nowacki W. On densest packings of ellipsoids. Zeitschrift für Kristallographie-Crystalline Materials. 1966; 123(1-6):401–421.

25. Koch E, Fischer W. Packings of ellipses and ellipsoids. Wiley Online Library; 2006.

26. Conway JH, Sloane NJA. Sphere packings, lattices and groups. vol. 290. Springer Science & Business Media; 2013.

27. Nowacki W. Symmetrie und physikalisch-chemische Eigenschaften krystallisierter Verbindungen. II. Die allgemeinen Bauprinzipien organischer Verbindungen. Helvetica Chimica Acta. 1943; 26(2):459–462. https://doi.org/10.1002/hlca.19430260210

28. Donev A, Stillinger FH, Chaikin P, Torquato S. Unusually dense crystal packings of ellipsoids. Physical Review Letters. 2004; 92(25):255506. https://doi.org/10.1103/PhysRevLett.92.255506 PMID: 15245027

29. Edelsbrunner H, Harer J. Computational Topology: An Introduction. American Mathematical Society, Providence, RI, USA; 2010.

30. Carlsson G. Topology and data. Bulletin of the American Mathematical Society. 2009; 46(2):255–308. https://doi.org/10.1090/S0273-0979-09-01249-X

31. Chazal F, Michel B. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. arXiv preprint arXiv:171004019. 2017.

32. Edelsbrunner H, Harer J. Persistent homology-a survey. Contemporary mathematics. 2008; 453:257–282.

33. Edelsbrunner H, Morozov D. Persistent homology: theory and practice. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2012.

34. Morozov D. Dionysus. Software available from http://wwwmrzvorg/software/dionysus. 2012.

35. Bender A, Glen RC. Molecular similarity: a key technique in molecular informatics. Organic & Biomolecular Chemistry. 2004; 2(22):3204–3218. https://doi.org/10.1039/b409813g PMID: 15534697

36. Wale N, Watson IA, Karypis G. Comparison of descriptor spaces for chemical compound retrieval and classification. Knowledge and Information Systems. 2008; 14(3):347–375. https://doi.org/10.1007/s10115-007-0103-5

37. Rogers D, Hahn M. Extended-connectivity fingerprints. Journal of Chemical Information and Modeling. 2010; 50(5):742–754. https://doi.org/10.1021/ci100050t PMID: 20426451

38. Morgan H. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of Chemical Documentation. 1965; 5(2):107–113. https://doi.org/10.1021/c160017a018

39. pfnet research. Chainer Chemistry; 2017.

40. Hoja J, Reilly AM, Tkatchenko A. First-principles modeling of molecular crystals: structures and stabilities, temperature and pressure. Wiley Interdisciplinary Reviews: Computational Molecular Science. 2017; 7(1):e1294.

41. Wu J, Pisula W, Müllen K. Graphenes as potential material for electronics. Chemical Reviews. 2007; 107(3):718–747. https://doi.org/10.1021/cr068010r PMID: 17291049

42. Silinsh EA. Organic molecular crystals: their electronic states. vol. 16. Springer Science & Business Media; 2012.

43. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials. 2016; 72(2):171–179. https://doi.org/10.1107/S2052520616003954

44. Desiraju GR, Gavezzotti A. Crystal structures of polynuclear aromatic hydrocarbons. Classification, rationalization and prediction from molecular structure. Acta Crystallographica Section B: Structural Science. 1989; 45(5):473–482. https://doi.org/10.1107/S0108768189003794

45. Ito D, Shirasawa R, Hattori S, Tomiya S, Tanaka G. Prediction of Molecular Packing Motifs in Organic Crystals by Neural Graph Fingerprints. In: International Conference on Neural Information Processing. Springer; 2018. p. 26–34.

46. Frisch M, Trucks G, Schlegel H, Scuseria G, Robb M, Cheeseman J, et al. Gaussian 16; 2016.

47. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

48. Pulido A, Chen L, Kaczorowski T, Holden D, Little MA, Chong SY, et al. Functional materials discovery using energy–structure–function maps. Nature. 2017; 543(7647):657–664. https://doi.org/10.1038/nature21419 PMID: 28329756