Applications of Persistent homology to materials science, and persistent homology software HomCloud

Ippei Obayashi

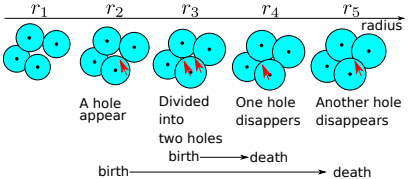
Okayama University

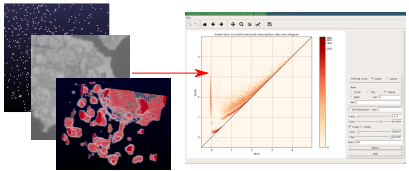
Aug. 7, 2025

Introduction to persistent homology

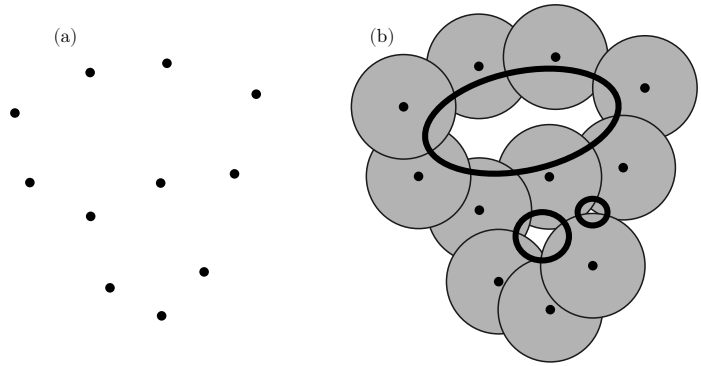
Persistent homology

- *Topological Data Analysis* (TDA)
  - Data analysis using the mathematical idea of topology
  - Quantitatively characterize the shape of data
    - ★ Connected components, rings, cavities,
- *Persistent homology* (PH): Mail tool of TDA
  - Using the idea of homology
  - PH gives good descriptors for the shape of data (called persistence diagram)
- Developed in the 21st century
  - Mathematical theory and algorithm
  - Software
  - Applications to materials science, life science, etc.





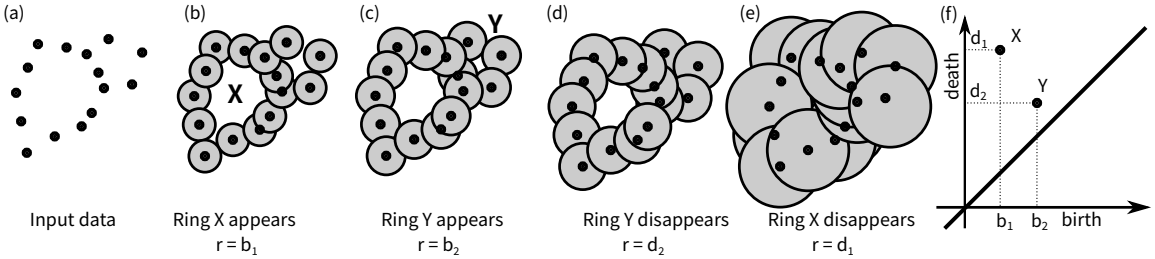
r-Ball model



- The input data is a set of points (a pointcloud)
- The pointcloud has no holes/rings, but it looks like some holes
- Put discs with radii r
- Three holes
- We can characterize the pointcloud by these holes

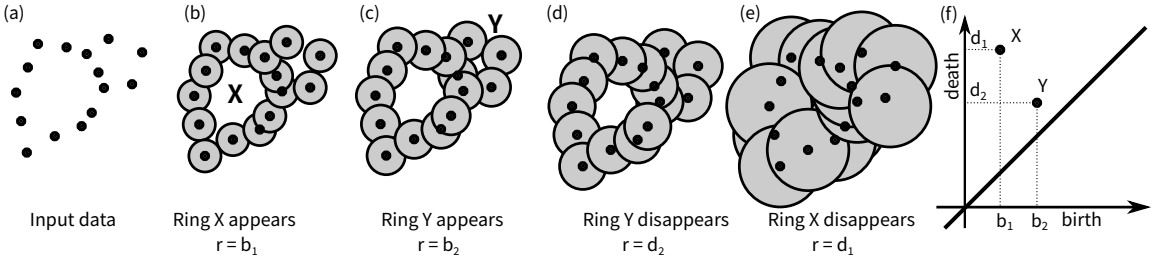
Filtration

The radius r becomes larger, and holes appear and disappear. The theory of persistent homology makes pairs of the appearance and disappearance of holes



Persistence diagram

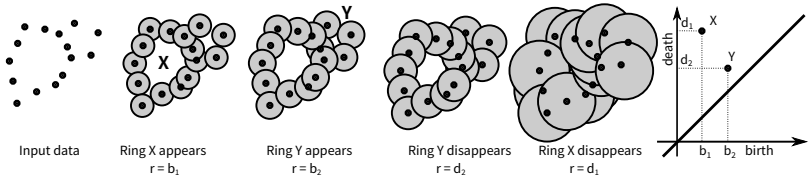
These pairs are called *birth-death pairs*, and the collection of birth-death pairs is called a *persistence diagram* (PD). The scatter plot (or 2D histogram plot) of the pairs is used to visualize PDs.



The persistence diagram gives a geometric summary of the data.

Inverse analysis

- In a PD, each birth-death pair corresponds to a ring or cavity
- It is very useful to identify such a ring/cavity
- Since there are many candidates, the best ring or cavity is selected using mathematical optimization.
  - Tightness of a homology generator is usually used as a criterion



Applications to materials research

Joint work with E. Minamitani, T. Shiga, and M. Kashiwagi [1]

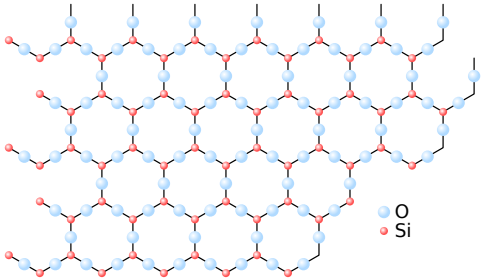
Outline of the research

- Aim: Understanding the relationship between nanostructures of amorphous silicon and thermal conductivity of the material
- Method: The thermal conductivity and persistence diagrams are calculated from a lot of atomic configurations of amorphous silicon by molecular dynamical simulations, and two statistical methods (ridge regression and principal component analysis) and an inverse analysis is applied to the diagrams to identify local structures that correlate with the thermal conductivity
- Results: Thermal conductivity changes when pentagons in the atomic configurations are deformed. We also successfully explained why the deformation changes the thermal conductivity from the viewpoint of physics, using vibrational modes on local structures

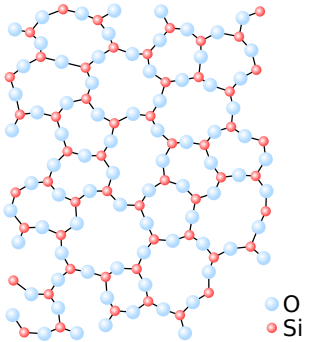
Background

Two types of solid materials: crystal and amorphous

- Crystal: periodic atomic structures
- Amorphous: No periodic structures

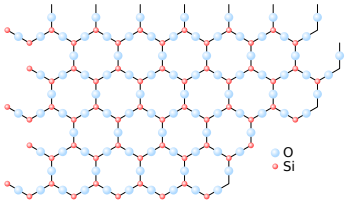


From Wikipedia: Solid

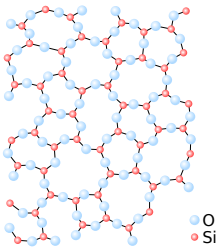


Short-range order, medium-range order, and long-range order

- Short-range order (SRO): Order on neighbor atoms
  - Numbers of adjacent atoms, bond lengths, and bond angles
  - Common for crystals, liquids, and amorphous
- Long-range order (LRO): Order on periodic (crystalline) structures
  - Only crystals
- Medium-range order (MRO): Order larger than short-range order
  - Common for crystals and amorphous, and liquids do not have this type of order
  - Clarifying medium-range order is an important problem on amorphous study



From Wikipedia: Solid



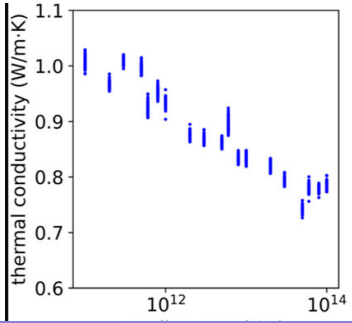
Thermal conductivity

Heat flows from a hot environment to a cold one. Heat flux, a flow of energy per unit area and per unit time, is described by Fourier's low:

$$-k\nabla T \tag{1}$$

where ∇T is the temperature gradient, and k is a constant determined by the material. k is called *thermal conductivity*.

- Usually, crystals have higher thermal conductivity than amorphous materials
- Amorphous materials have different thermal conductivity depending on their structures
  - Amorphous structures are calculated in molecular dynamical simulations by rapid cooling of hot materials
  - The cooling rate and thermal conductivity are correlated (the figure)
  - The cooling rate and medium-range order are also considered to be correlated
  - Therefore, we have a *hypothesis* that medium-range order determines thermal conductivity



Problem

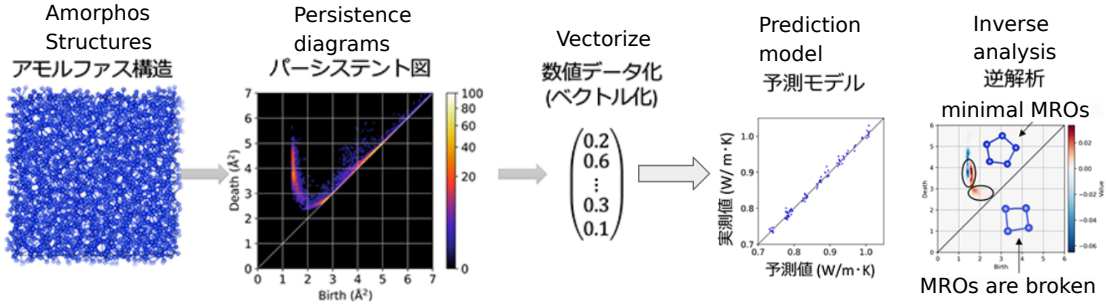
How do we characterize medium-range orders quantitatively?

- Descriptors suitable for our purpose
- Recently persistent homology has successfully described amorphous structures like silica glass and metallic glass

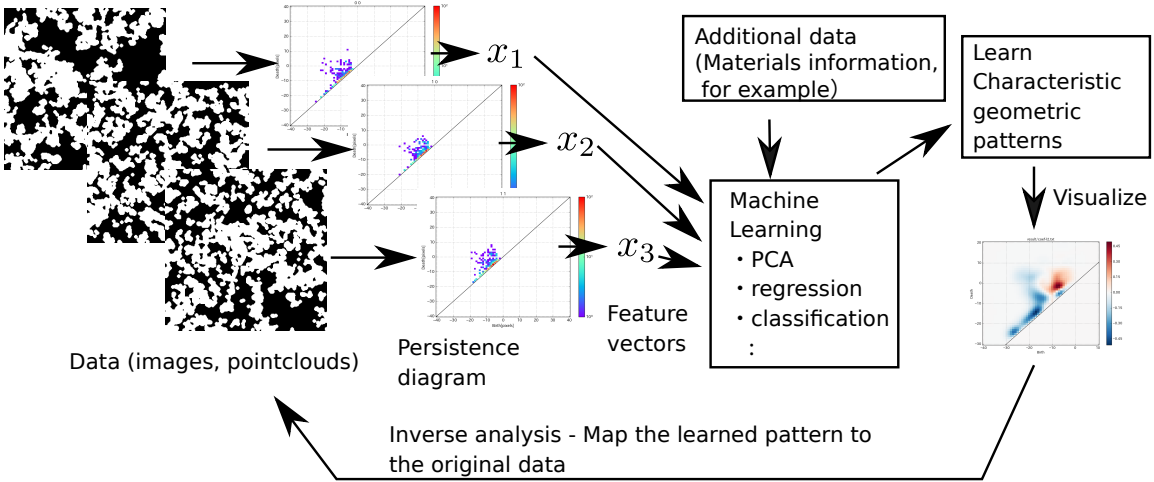
We tried to apply PH for our purpose.

Method

The thermal conductivity and persistence diagrams are calculated from a lot of atomic configurations of amorphous silicon by molecular dynamical simulations, and two statistical methods and the inverse analysis is applied to the 1st diagrams to identify local structures that correlate with the thermal conductivity.



Machine learning framework



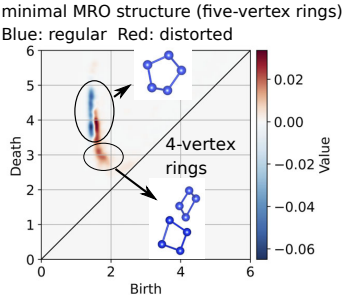
Obayashi et al. (2018)[2]

Machine learning concepts

- PH can summarize the *geometric* structure of the data quantitatively
- Machine learning can find *characteristic patterns* from the data
- The combination of these two concepts, we can extract *characteristic geometric patterns* from the data
- By combining *linear machine learning model*, *histogram-based vectorization* of persistence diagrams, and *inverse analysis*, we realize *explainable* PH machine learning!

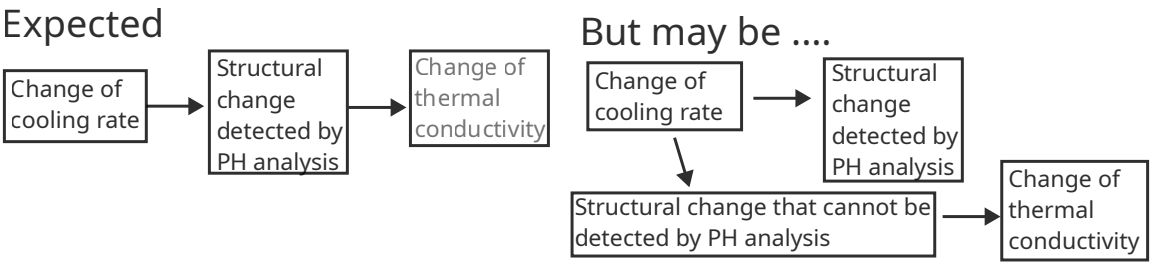
Results

Thermal conductivity changes when pentagons in the atomic configurations are deformed.



- The blue area positively contributes to thermal conductivity, and the red area negatively contributes to thermal conductivity
- The deformation of five-vertex rings are correlated with thermal conductivity

In the PH analysis, we only found the correlation between thermal conductivity and 5-vertex / 4-vertex rings, and both of the following two graphs are possible:



We need to identify the cause of the relationship through:

- theoretical analysis
- further experiments

We successfully explained why the deformation changes the thermal conductivity from the viewpoint of physics, using vibrational modes on local structures

- We estimated the influence of local structures on thermal conductivity using vibrational modes from Allen–Feldman theory in physics
- We found that the deformation of five-vertex rings makes the thermal conductivity lower

Software

Which software do we use? You can use any software you like, but our software, HomCloud, is a good option!

- HomCloud has powerful inverse analysis tools, such as optimal and stable volumes, PH trees, and optimal 1-cycles
- HomCloud has vectorization functionality by persistence image and persistence codebook
- HomCloud has a useful Python interface, and you can use HomCloud with Python's rich machine-learning ecosystem
- All necessary components for our PH+ML framework are available and you can combine these components flexibly
- HomCloud is available on Linux, Apple Silicon Mac, Windows, and Google Colaboratory

`https://homcloud.dev`

Development

For better software, we are implementing the following practices:

- Dogfooding
  - The developer always uses HomCloud for PH data analysis
  - Software quality can be monitored at all times
  - Usability is always tested by the developers
- Continuous integration
  - The software is automatically built and tested on Linux, Windows, and Mac every time the code is changed

How to learn HomCloud

HomCloud has the Python interface and CLI, but the Python interface is recommended. If you want to try HomCloud, it is best to start with Google Colaboratory.

- Only a web browser and a Google account
- Lightweight installation process

HomCloud provides tutorials at `https://homcloud.dev/tutorials.en.html`, starting with "3D Point Cloud Data Analysis". After completing that tutorial, you should proceed to other tutorials that interest you.

HomCloud Documentation

HomCloud provides two types of documentation on the website.

- Tutorials
  - Each tutorial provides the complete set of introductory knowledge about persistent homology data analysis
  - You can choose tutorials on topics that interest you.
    - ★ 3D point cloud analysis, binary or grayscale image analysis
    - ★ More advanced topics: atomic configurations, machine learning
- Python API documentation
  - Detailed descriptions of each class and method
  - If the API document is insufficient, you can use the source code of HomCloud itself

How to cite HomCloud

You can use HomCloud freely since it is open source software.

Please cite the following review paper [3] about persistent homology and HomCloud. You may also refer to the URL of this website, `https://homcloud.dev/index.en.html`, with HomCloud's version.

I. Obayashi, T. Nakamura, and Y. Hiraoka. "Persistent Homology Analysis for Materials Research and Persistent Homology Software: HomCloud". In: J. Phys. Soc. Jpn. 91.9 (2022), p. 091013. doi: 10.7566/JPSJ.91.091013.

Future of HomCloud

Now we plan:

- Continuous code refinement
- Updating documents
  - Recently, the tutorials have been reviewed and refreshed
  - The next step is to prepare code snippets
- Publicly publishing the source code repository
  - Now the source code of HomCloud is available by downloading the .tar.gz file, but the Git repository is not available.

More ambitious ideas for developing HomCloud:

- Parallel computation for optimal or stable volumes
  - Since they require significantly more cost than computing PDs.
- Non-rectangular unit cell for periodic boundary condition for 3d alpha filtrations
  - CGAL supports a rectangular unit cell, and HomCloud uses the feature
  - CGAL does not support a non-rectangular parallelepiped unit cell
  - It is a desirable feature, but difficult to implement
- Integration with connected PD [4] software, one of extensions of PDs using multi-parameter persistent homology

Acknowledgments

Collaborators:
Y. Hiraoka, M. Kimura, E. Minamitani, T. Shiga, M. Kashiwagi

Funding Support:
JSPS Kakenhi Grant number 16K17638, 19H00834, 19KK0068, 20H05884, 22H05106,
JST PRESTO Grant Number JPMJPR1923, JST CREST Grant Number JPMJCR15D3.

Wrap up

- Applications of persistent homology to material research
- Our software, HomCloud, is available for PH data analysis

References

[1] Emi Minamitani, Takuma Shiga, Makoto Kashiwagi, and Ippei Obayashi. Topological descriptor of thermal conductivity in amorphous Si. Journal of Chemical Physics **156**, 244502 (2022)

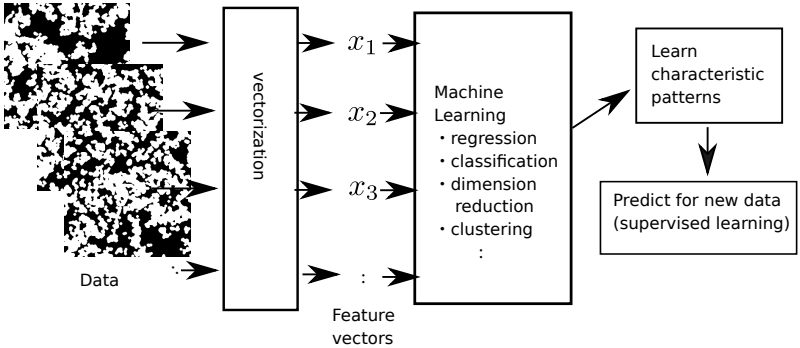
[2] Ippei Obayashi, Yasuaki Hiraoka, and Masao Kimura. Persistence diagrams with linear machine learning models. Journal on Applied and Computational Topology **1**, 3–4, 421–449, (2018).

[3] I. Obayashi, T. Nakamura, and Y. Hiraoka. Persistent Homology Analysis for Materials Research and Persistent Homology Software: HomCloud. J. Phys. Soc. Jpn. **91**, 091013 (2022)

[4] Y. Hiraoka, K. Nakashima, I. Obayashi, and C. Xu. Refinement of Interval Approximations for Fully Commutative Quivers. Arxiv preprint (2023). `https://arxiv.org/abs/2310.03649`

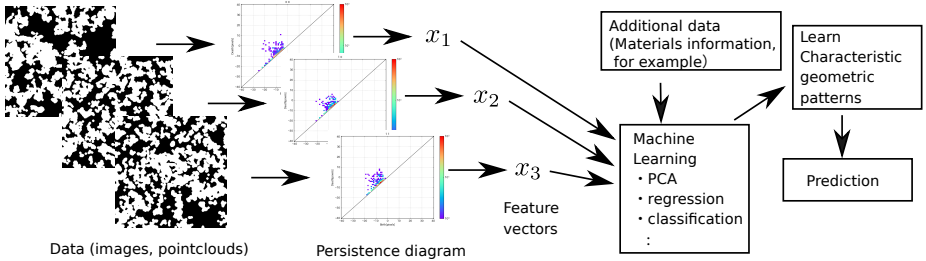
Appendix page

Machine leaning



- Standard ML framework, each data is converted into a vector, and an ML method is applied to find characteristic patterns in the data
- In the case of supervised learning, the patterns are used to predict
- For good performance of the learning, we need to construct a good vector
  - Neural networks can learn how to vectorize from data

Standard framework of the combination of persistent homology and machine learning



- We need to convert PDs to vectors to apply ML to PH
- How to vectorize is an important problem

Many vectorization methods are proposed:

- Persistence landscape
- PSSK
- PWGK
- Persistence image
- Persistence codebook
- Sliced Wasserstein Kernel
- Persistence Fisher Kernel
- :

In this talk, we use Persistence Image (PI) for vectorization[2]. We use linear models for machine learning to enhance the interpretability of the learned result.

Persistence Image[2]

- The histogram of a PD is regarded as a vector
- Birth death pairs near the diagonal are less important, and Gaussian kernel density estimation is used.
- The following function on a plane is discretized
  - For $D_q = \{(b_k, d_k)\}_{k=1}^{M}$, we define a function ρ on (x, y) plane as follows:

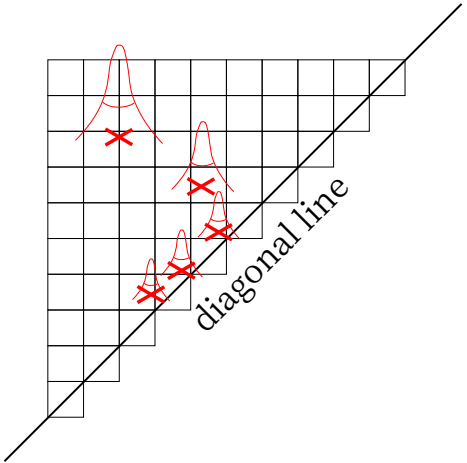
$$\rho(x, y) = \sum_{k=1}^{M} w(b_k, d_k) N(x, y \mid b_k, d_k)$$

$$w(b, d) = \arctan(C(d - b)^p)$$

$$N(x, y \mid b_k, d_k) = \exp(-\frac{(b_k - x)^2 + (d_k - y)^2}{2\sigma^2})$$

  - $C, p, \sigma > 0$ are parameters
  - ρ is discretized by grids, We can effectively compute the vector by Gaussian filter

Schematic figure of persistence image



Machine learning models

- Linear models are used
  - Linear regression
  - Logistic regression
  - Principal Component Analysis (PCA)
  - Non-negative Matrix Factorization (NMF)
  - :
- Easy to interpret the learned result

Linear regression

- Basic regression model
  - Supervised learning
- Data: $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^k \times \mathbb{R}$
- Goal: Find the relationship between x and y
- Model: We assume the following relationship between x and y

$$y = a \cdot x + b + \epsilon$$

ϵ means a random noise. Proper $a \in \mathbb{R}^k, b \in \mathbb{R}$ are estimated from give data (this is "learning")

$$y = a \cdot x + b + \epsilon$$

To combine linear regression with PH, PI vector is used as x

$$y = a \cdot \rho(D_q(F(X))) + b + \epsilon$$

- X: Input data for PH (a pointcloud or an image)
- $F(X)$: The filtration built from X
- $D_q(\cdot)$: q-dim PD
- $\rho(\cdot)$: The PI vector from $\cdot$
- a, b: The parameters of linear regression

Then the inner product in the regression model is approximately L^2 inner product

$$y = \sum_{(b_i,d_i)} \int a(u,v)w(b_i,d_i)N(u,v \mid b_i,d_i)dudv + b$$

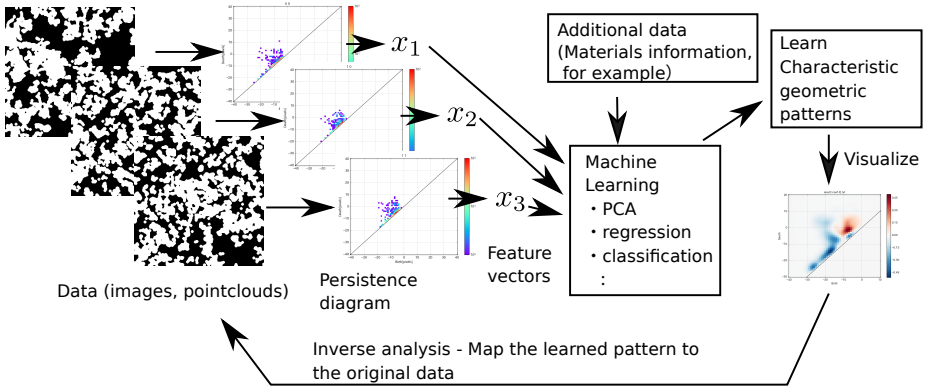
$a(u,v)$ is the learned vector. Then,

$$\int a(u,v)w(b_i,d_i)N(u,v \mid b_i,d_i)dudv$$

means the "importance" or "contribution" of each birth-death pair (b_i,d_i)

We can visualize $a(u,v)$ in the form of PDs(dual persistence diagram).

By applying *inverse analysis* to those important pairs, we can visualize the shape corresponding to these important pairs.



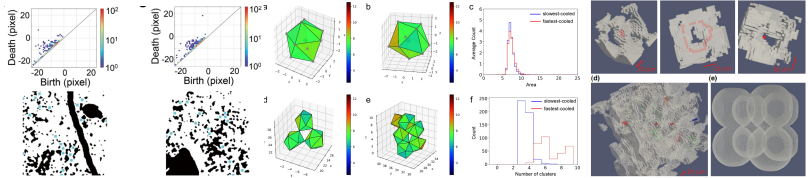
Some other materials applications

Target	Data type	ML models	Inverse analysis method
iron-ore sinters [3]	2D images	PCA & Lasso	birth or death pixels
metallic glass [4]	pointclouds	PCA & RR	optimal volume
Si amorphous [5]	pointclouds	PCA & RR	optimal volume
iron-ore sinters [6]	3D images	NMF	stable volume & optimal 1-cycle

RR=Ridge regression

The target materials, data types, machine learning models, and inverse analysis methods are different, but the fundamental ML+PH framework is common.

[3] [4] [6]



Selection of machine learning method

In this study, we used linear regression and principal component analysis as machine learning models. To apply the framework to your own data, you may wonder which ML model is better.

- I recommend PCA as a first choice
  - Since PCA is robust against data fluctuation compared to other models
- If the PCA result looks good, you can try various linear models such as logistic regression, linear SVM, linear regression, and nonnegative matrix factorization
  - Regularizaiton is important
- If you are not satisfied with the performance of the linear models, it is worthwhile to work with nonlinear methods such as gradient boosting tree models or neural network models
  - To identify which birth-death pairs are important, I recommend feature selection methods or explainable AI methods such as SHAP

Selection of vectorization method

Is persistence image the best way to vectorize?

- I think that histogram-based methods are better because of the interpretability, and persistence image is one option
- In fact, simple histogram-based methods (without Gaussian diffusion and weighting based on the distance to the diagonal) sometimes give sufficient prediction accuracy
- Adaptive mesh is potentially a good choice
  - Some of persistence codebook methods