Topological Data Analysis on Materials Science

Yasu Hiraoka

WPI-AIMR, Tohoku University

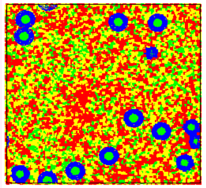
Supported by

JST CREST
SIP Structural Materials for Innovation
JST Innovation Hub MI^2I, NIMS
NEDO

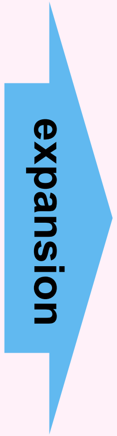
Materials TDA

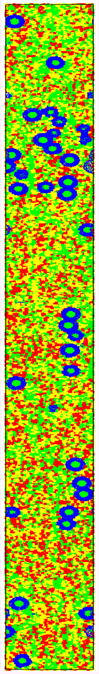
Supported by AIMR, CREST, SIP, MI^2I, NEDO

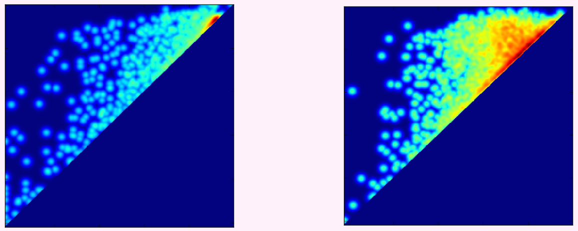
Polymer



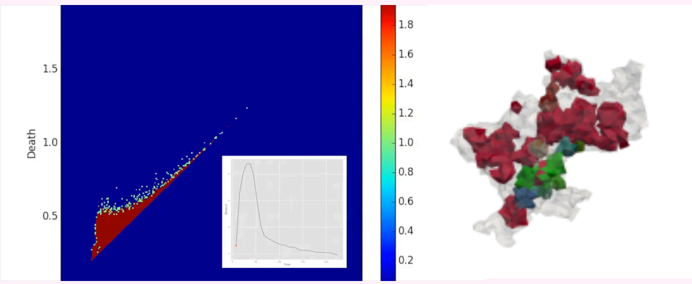
Atomic Force Microscopy image (by Nakajima)







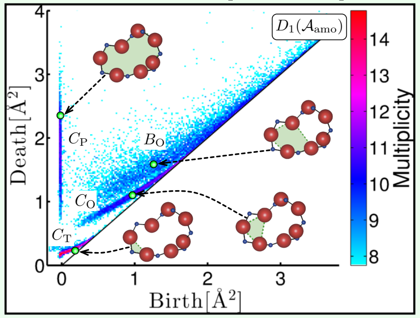
craze formation

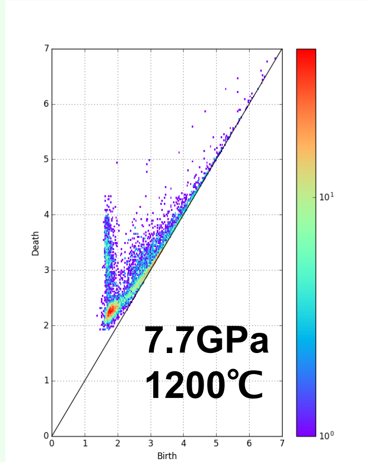


PRE (2017)

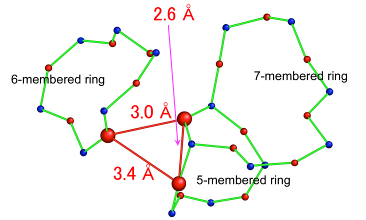
Glass

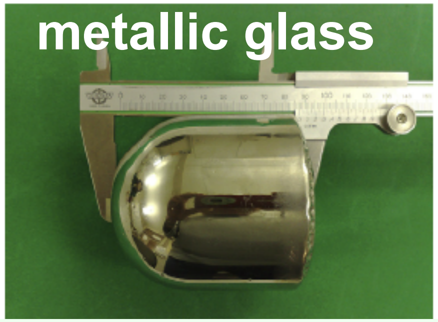
PNAS (2016)

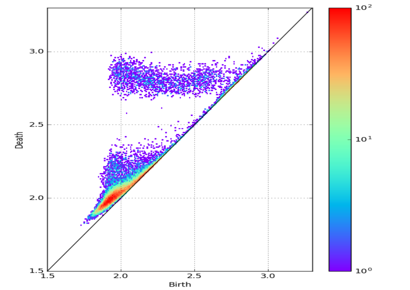




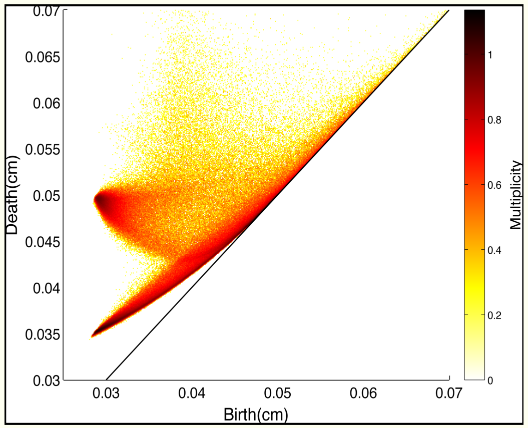
densified silica



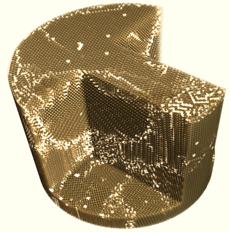




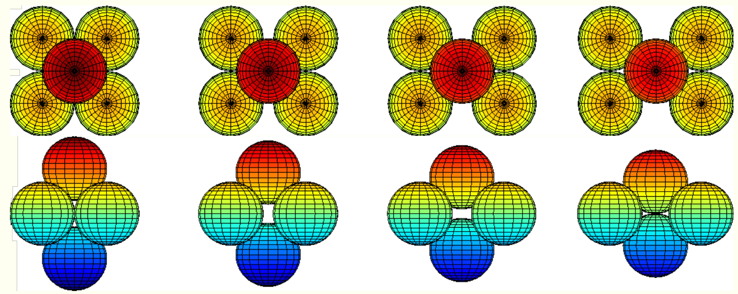
Grain



Nature Communications (2017)

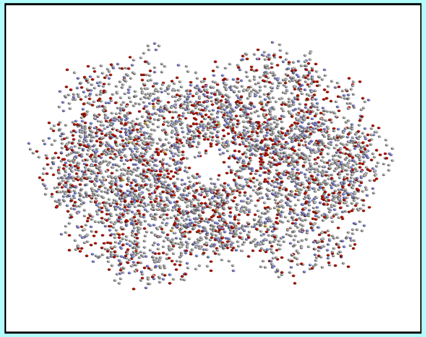


deformation of octa.



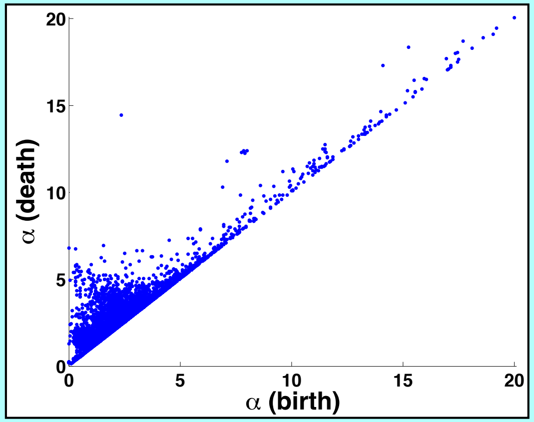
New math: Persistent homology

Input data → Persistent Homology → Shape of data



Atomic configuration of hemoglobin

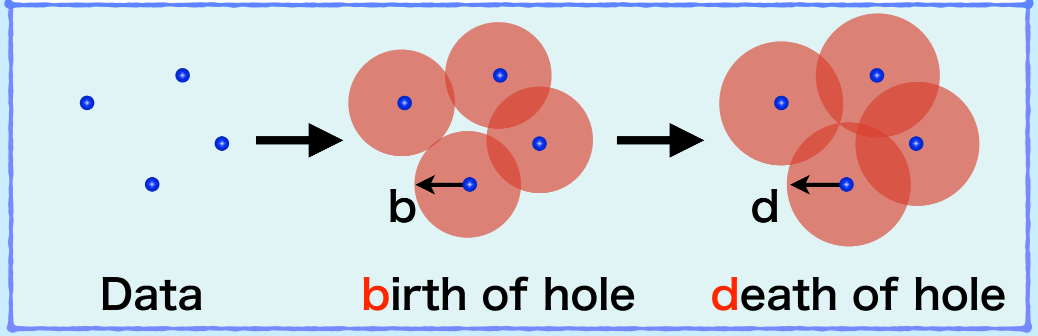
- characterize holes in data
- describe number, size, and shapes
- multi-scale analysis



Persistence Diagram (PD)

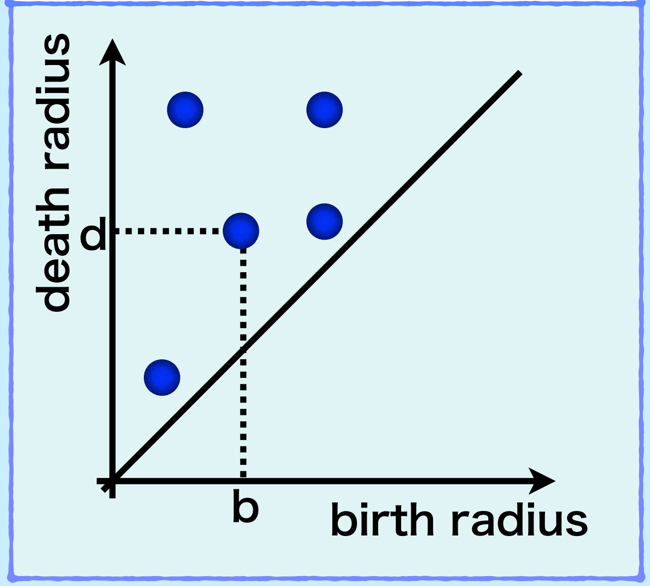
Persistence diagram of point cloud

Fattening (filtration)



inverse

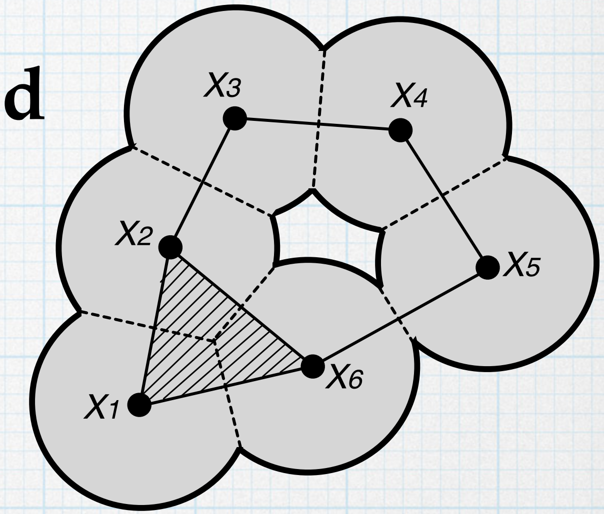
Persistence diagram



- each point (called generator) in PD expresses a hole in data
- birth & death axes measure shapes of holes
- points close to diagonal are noisy
- points away from diagonal are robust

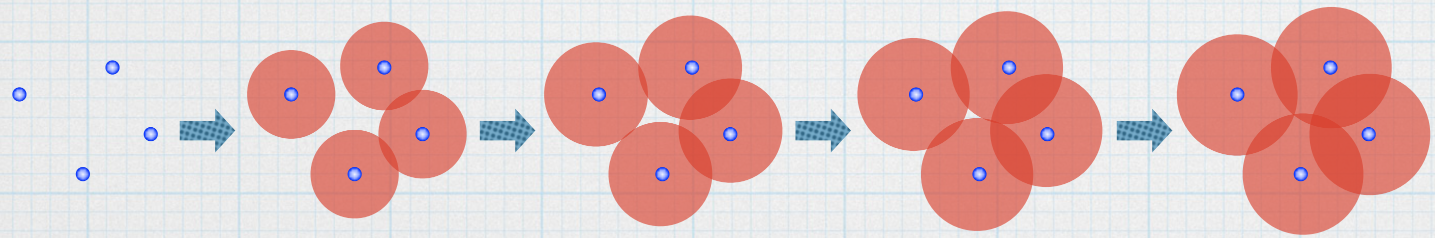
Note: 2D histogram uncovers further geometry

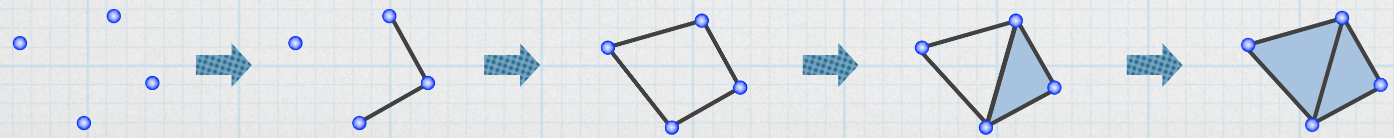
Alpha filtration



- $X = \{x_i \in \mathbf{R}^m \mid i = 1, \ldots, n\}$ **: point cloud**
- $\mathbf{R}^m = \cup_i V_i$ **: Voronoi decomp.**
- $\cup_i B_i(r) = \cup_i (B_i(r) \cap V_i)$
- **Alpha shape** $\mathcal{A}(X, r)$ **: dual of** $\{B_i(r) \cap V_i \mid i = 1, \ldots, n\}$ **(simplicial complex)**
- **Nerve theorem:** $\cup_i B_i(r) \simeq \mathcal{A}(X, r)$
- $\mathcal{A}(X, r) \subset \mathcal{A}(X, s)$ **for** $r < s$

easy to analyze by computers

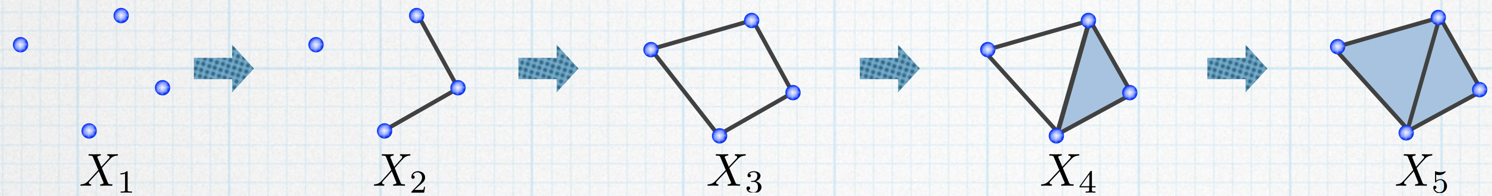




filtration:
changing resolution

Edelsbrunner, Letscher, Zomorodian, Carlsson, de Silva

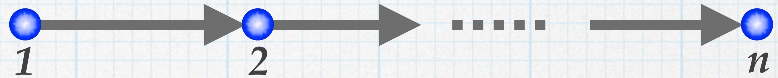
Persistent homology, diagram



$H_1(\mathcal{X}) \simeq I[3,4]$

- **filtration** $\mathcal{X} : X_1 \subset X_2 \subset \cdots \subset X_n$
- **persistent homology** $H_\ell(\mathcal{X}) : H_\ell(X_1) \to H_\ell(X_2) \to \cdots \to H_\ell(X_n)$

representations on A_n



- **interval decomposition (Gabriel Thm, fin.gen. PID module)**

$$H_\ell(\mathcal{X}) \simeq \bigoplus_{i=1}^{s} I[b_i, d_i]$$

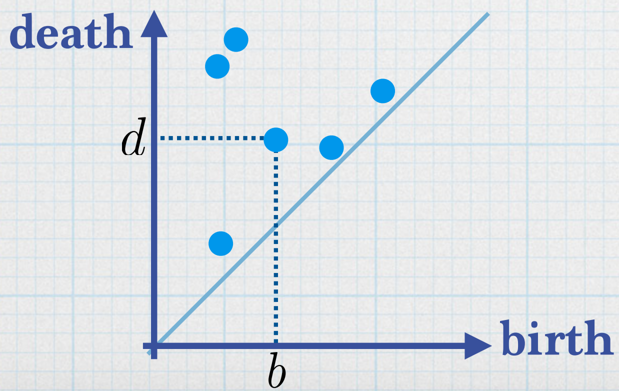
d - b : lifetime (or persistence)

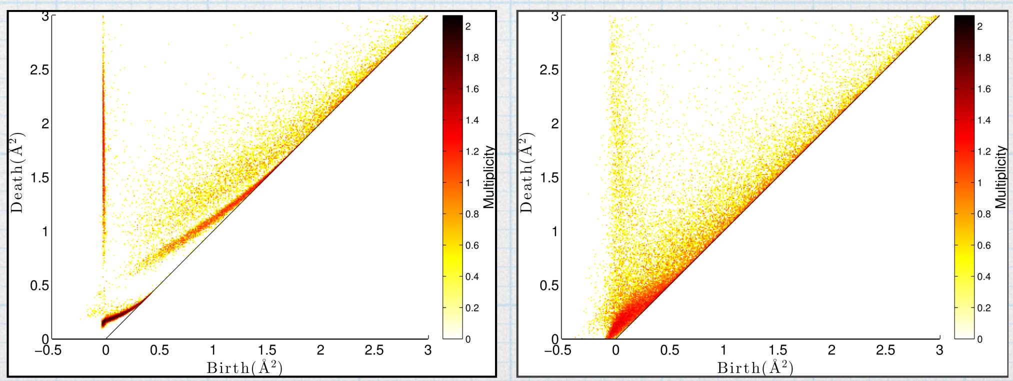
$$I[b,d] : 0 \to \cdots \to 0 \to K \to \cdots \to K \to 0 \to \cdots \to 0$$

at X_b **at** X_d

Each interval represents birth & death of a topological feature

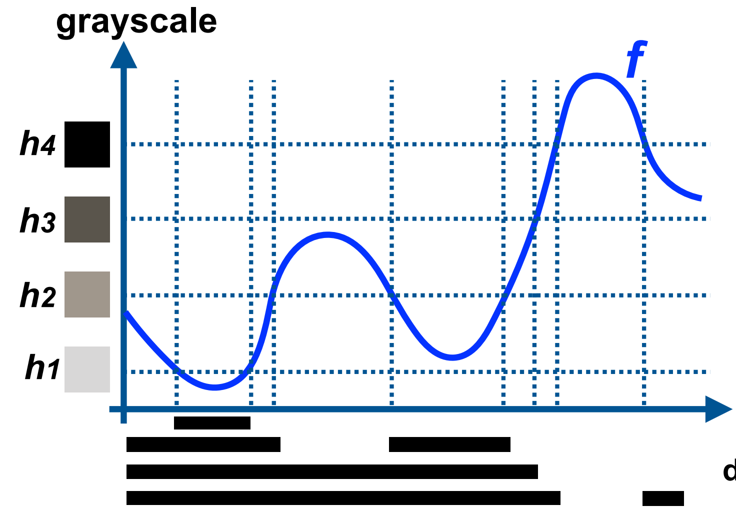
- **persistence diagram** $D_k(\mathcal{X}) = \{(b_i, d_i) \in \mathbb{R}^2_{\geq 0} : i = 1, \ldots, p\}$

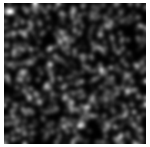




Persistent homology of digital image

1. Grayscale persistence

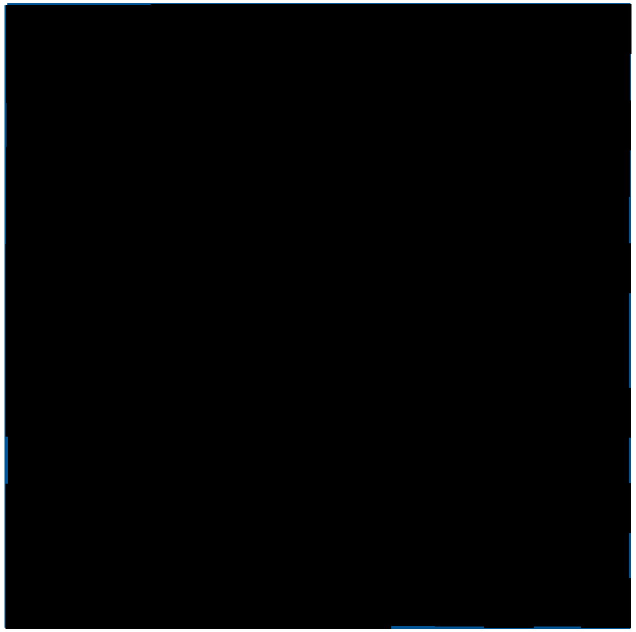


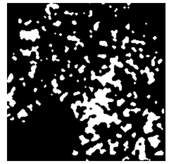


gray scale digital image

- **sub-level set** $X_h := \{x \in X \mid f(x) \leq h\}$
- **fattening** $X_{h_1} \subset X_{h_2} \subset \cdots \subset X_{h_T}$ **by** $h_1 \leq h_2 \leq \cdots \leq h_T$

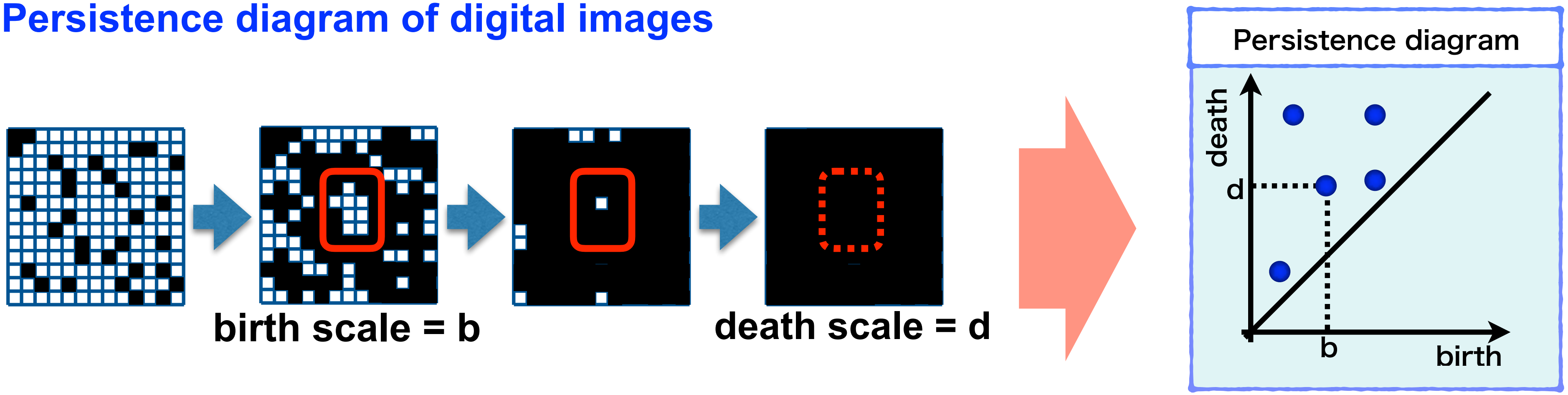
2. Spatial persistence





black-white image

Persistence diagram of digital images

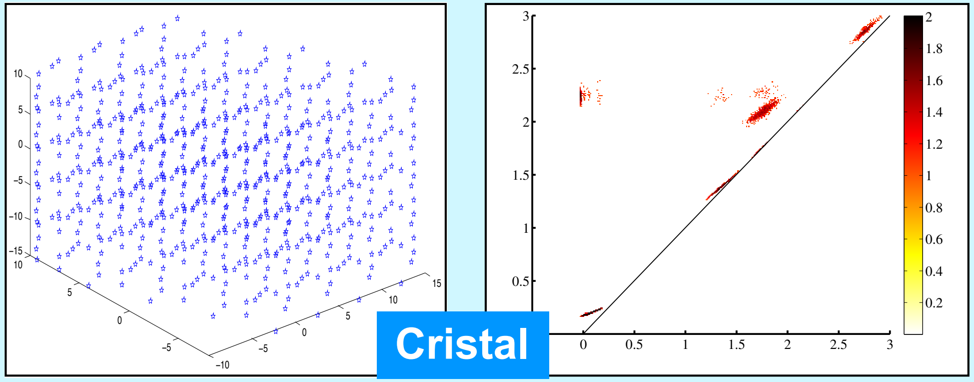


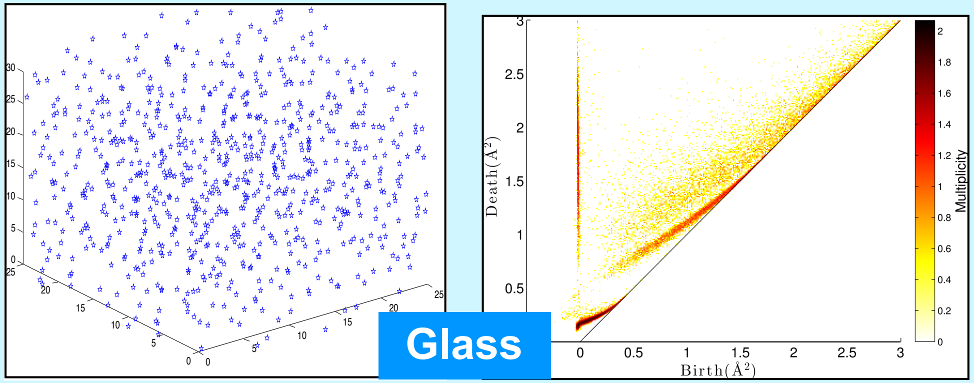
Characterize grayscale/spatial persistent holes in images

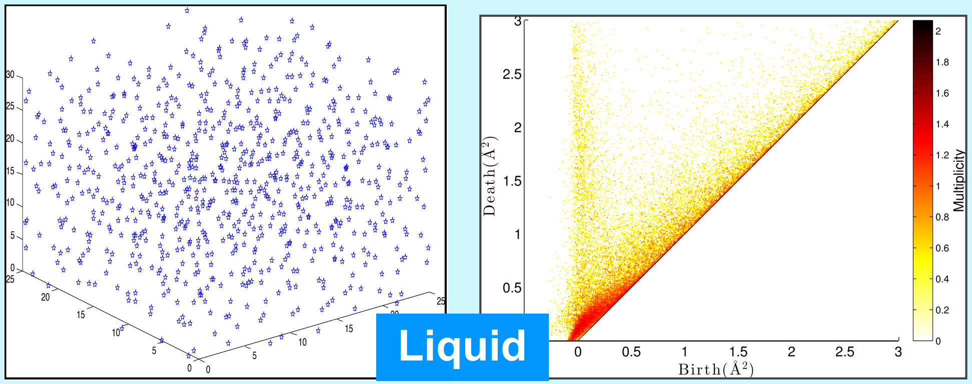
Hierarchical Structural Analysis of Silica Glass

with Nakamura, Hirata, Escolar, Matsue, Nishiura
PNAS (2016) CREST TDA, SIP

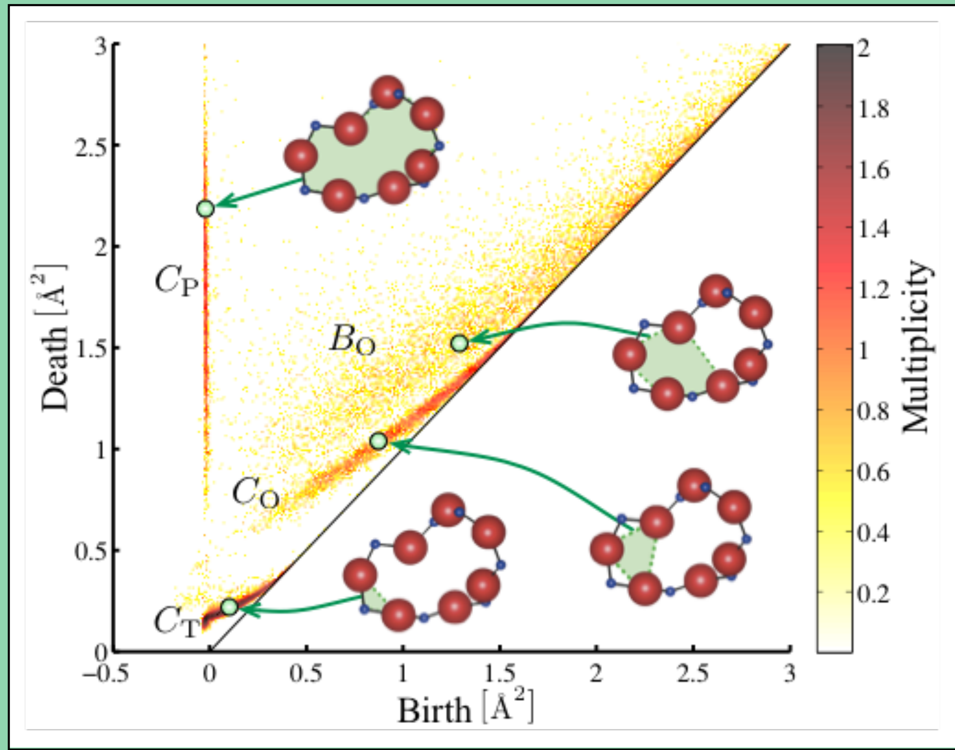
MD and PD_1







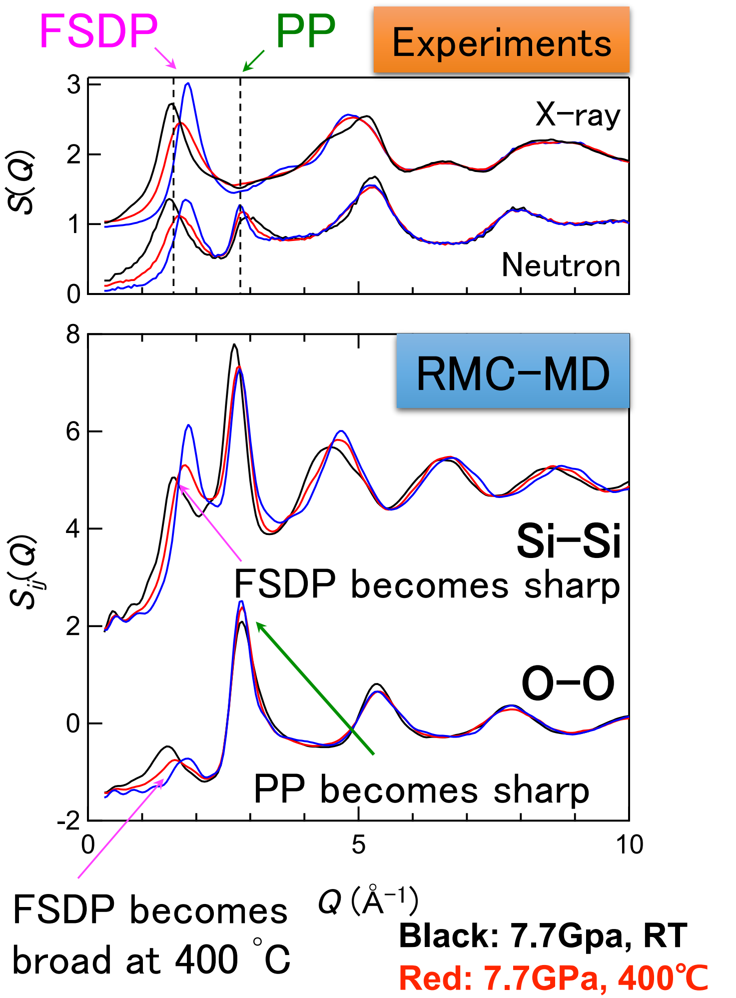
Inverse Analysis



- Glass contains curves in PD
- Curves express geometric constraints (orders) of atomic configurations
- Inverse analysis reveals hierarchical ring structures
- PD multi-scale analysis characterizes inter-tetrahedral O-O orders (curve Co)
- universal tool for structural analysis

Densified silica glass in high pressure and temperature

with Kohara (NIMS), Hirata, Obayashi (AIMR)   MI^2I (Innovation Hub), CREST TDA

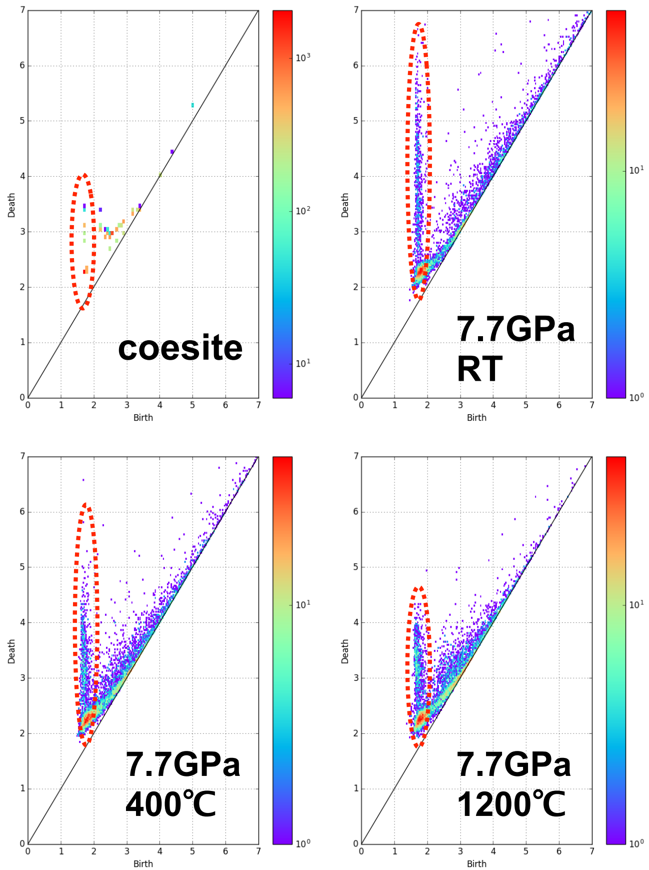


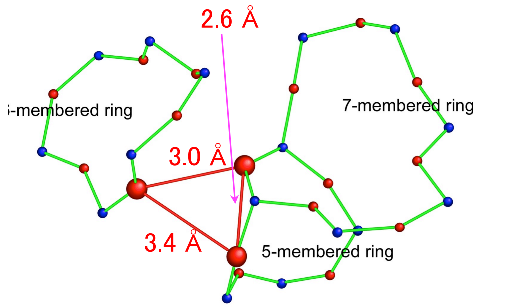
Black: 7.7Gpa, RT
Red: 7.7GPa, 400℃
Blue: 7.7GPa, 1200℃

- PP of O-O correlation becomes sharp with increasing temperature
- conventional methods could not explain this behavior

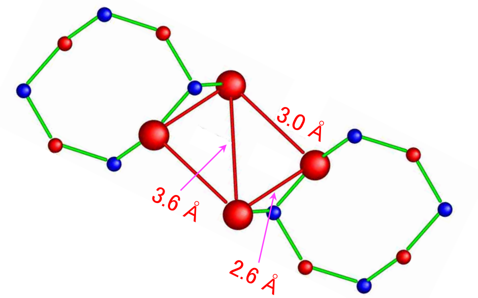
➡ what is the geometric origin?

O-PD1





inter-tetrahedral O-O ordering in 7.7GPa, 1200℃



ordering of oxygen in coesite

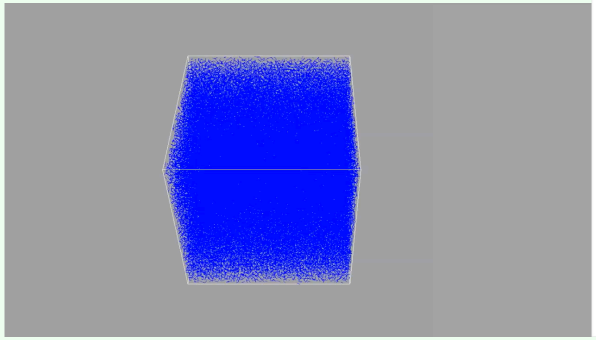
- **PDs become sharper like PP, and show the increase of packings of oxygens at high temp.**
- **Oxygen PDs ascribe for the first time O-O ordering between different SiO4 tetrahedra to PP**
- **The geometric origin of PP ordering is coesite-like rings**

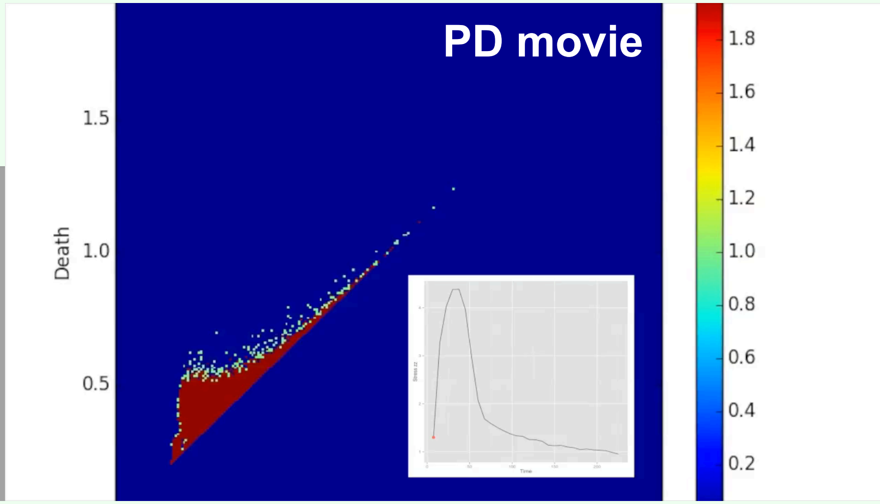
Craze formation of polymers

with Ichinomiya, Obayashi PRE (2017) SIP, NEDO

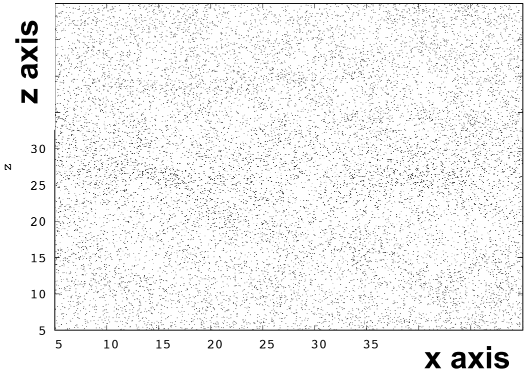
Kremer-Grest model

uniaxial deformation

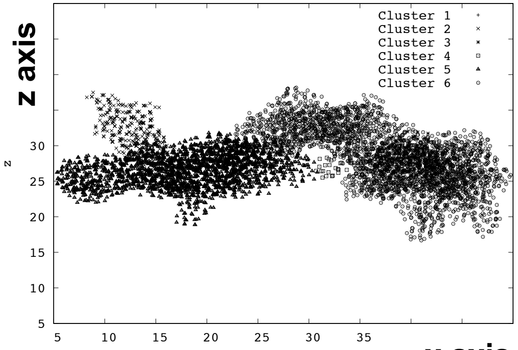




craze position

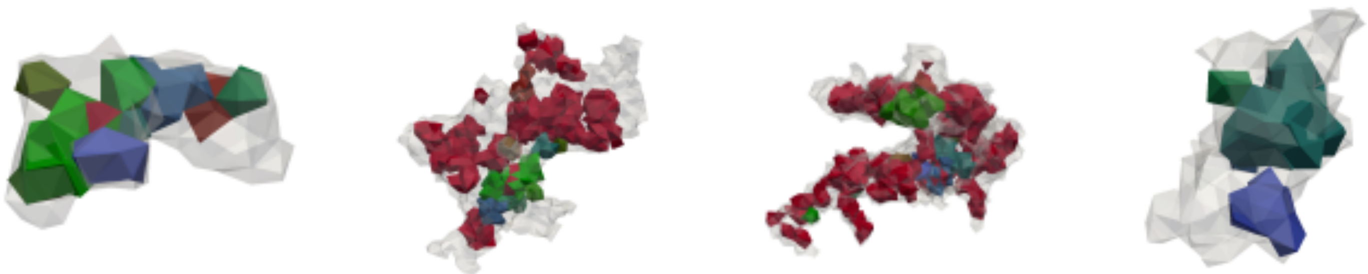


Voronoi volume (conventional)



PD method

void coalescence during craze formation



- gray voids are large voids observed after yielding
- color voids are initial micro voids generating large voids

- detect large voids from PD movie as generators with large death values
- explore initial config. of large voids by reversing time with inverse PD method
- large voids are generated by coalescence of micro voids (void percolation)

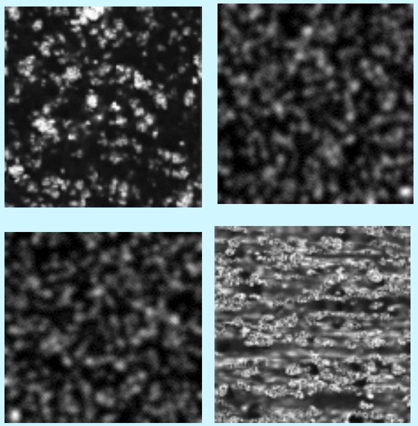
Statistical inverse analysis on persistence diagram

with Obayashi (AIMR) arXiv:1706.10082 CREST TDA, SIP, NEDO, MI^2

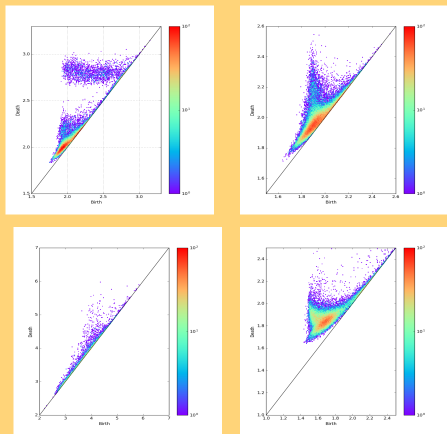
Background

- **PDs are good descriptors for disordered systems**
- **Want to extract statistical features encoded in dataset of PDs**
- **Vectorization of PDs are necessary for applying machine learnings (persistence landscape, persistence image, PSSK, PWGK, etc)**
- **Want to study the original data space (inverse problems)**

Dataset of inputs



Dataset of PDs



Machine learning

- PCA
- Regression
- Anomaly detection
- Time series
- etc

Study linear machine learning models based on persistence diagrams

Vectorization: persistence image
Linear ML: Logistic regression, Linear regression (LASSO/RIDGE)

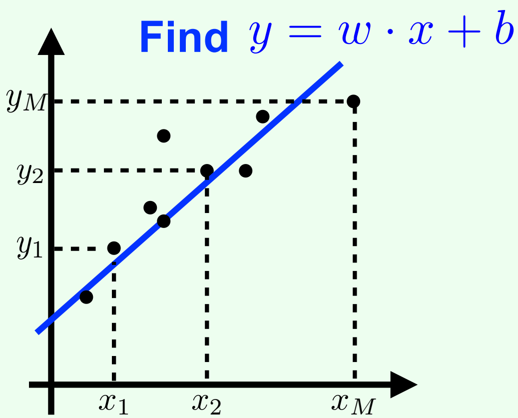
Linear regression of persistent homology

with Obayashi (AIMR) arXiv:1706.10082

Linear regression:

Given a training set $\{(x_i, y_i)\colon x_i \in \mathbf{R}^n, y_i \in \mathbf{R}\}_{i=1}^{M}$**,**
find optimal $w \in \mathbf{R}^n$ **and** $b \in \mathbf{R}$ **for the model**

$$y = w \cdot x + b + (\text{noise})$$



⟺ **find the minimizer**

$$E(w, b) = \frac{1}{2M} \sum_{i=1}^{M} (w \cdot x_i + b - y_i)^2 + \lambda R(w)$$

- **explanatory variable** $x \in \mathbf{R}^n$**: (vectorized) persistence diagram**
- **response variable** $y \in \mathbf{R}$**: conductivity, elasticity, crack area, etc**
- **Learned vector** w **can be expressed by PD (called learned PD)**

➡ **showing relevant generators in PDs to the response variable**

➡ **inverse of those generators explicitly shows relevant geometric features**

- **Suppress overfitting:**

LASSO PD: $R(w) = ||w||_1$
(sparse PD analysis)

RIDGE PD: $R(w) = \frac{1}{2}||w||_2^2$
(nice math property)

Logistic regression of persistent homology

Logistic regression:

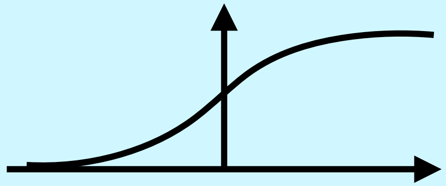
Given a training set $\{(x_i, y_i) \colon x_i \in \mathbf{R}^n, y_i \in \{0,1\}\}_{i=1}^{M}$ **,**

find optimal $w \in \mathbf{R}^n$ **and** $b \in \mathbf{R}$ **for the model**

$$P(y=1 \mid w, b) = g(w \cdot x + b),$$

$$P(y=0 \mid w, b) = 1 - P(y=1 \mid w, b) = g(-w \cdot x - b),$$

$$g(z) = 1/(e^{-z} + 1)$$



⟺ **find the minimizer**

$$L(w,b) = -\frac{1}{M}\sum_{i=1}^{M}\{y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\} + \lambda R(w)$$

$$\hat{y}_i = g(w \cdot x_i + b)$$

- **explanatory variable** $x \in \mathbf{R}^n$ **: (vectorized) persistence diagram**
- **response variable** $y \in \{0,1\}$ **: (binary) classification**
- **Learned vector** w **can be expressed by PD (called learned PD)**

➡ **generators in PDs with its inverse identify the relevant geometric features for classification**

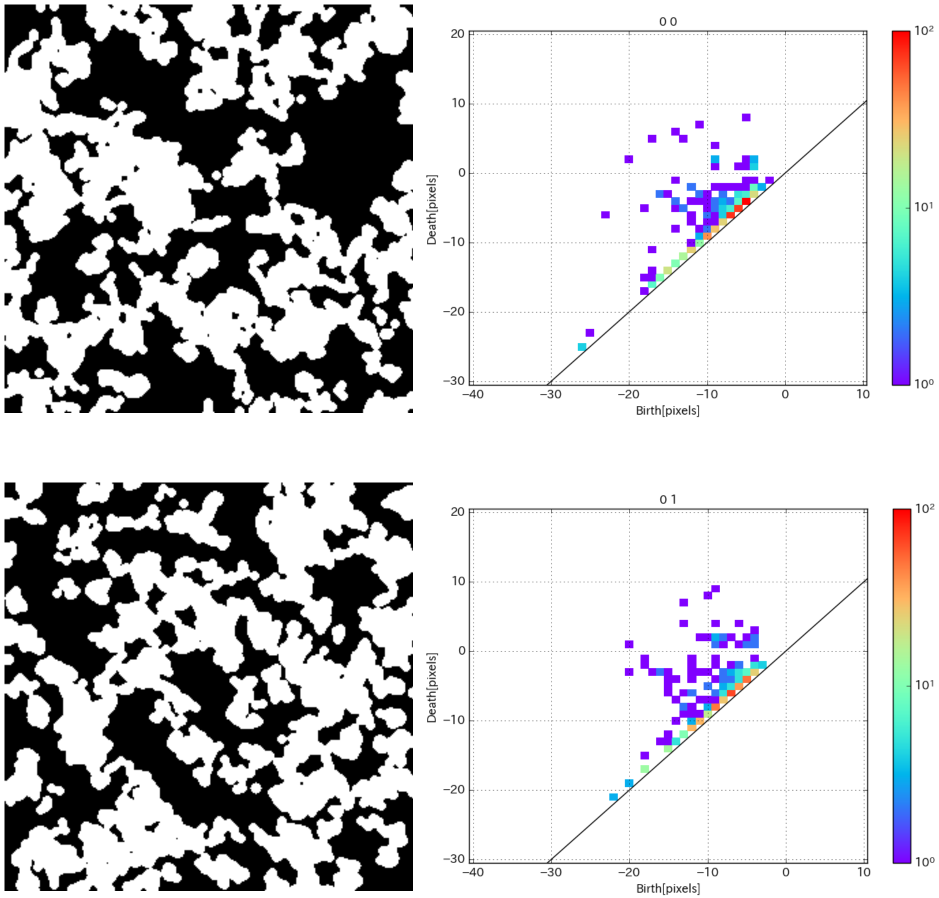
- **Suppress overfitting:**

LASSO PD: $R(w) = ||w||_1$
(sparse PD analysis)

RIDGE PD: $R(w) = \frac{1}{2}||w||_2^2$
(nice math property)

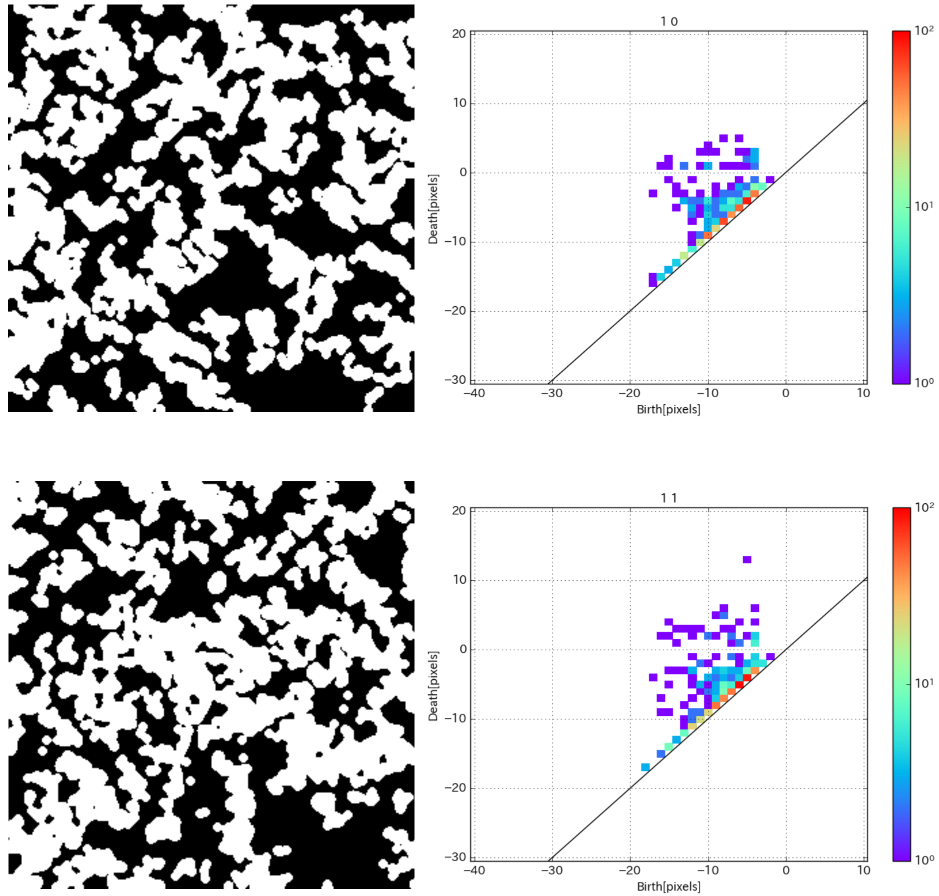
Performance of RIDGE logistic regressions: Easy example

Model A (200 trainings, 100 tests)



$y = 0$

Model B (200 trainings, 100 tests)



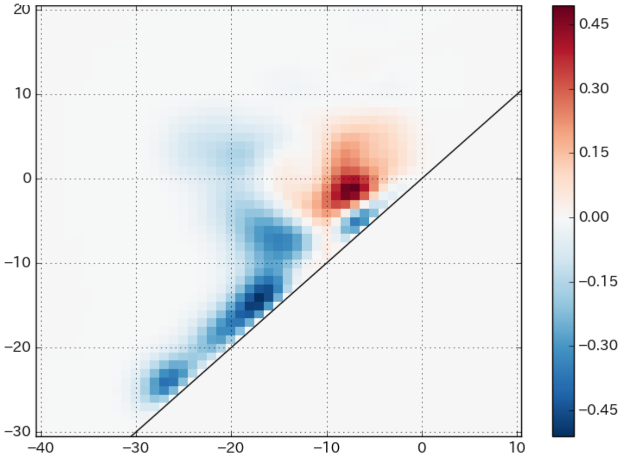
$y = 1$

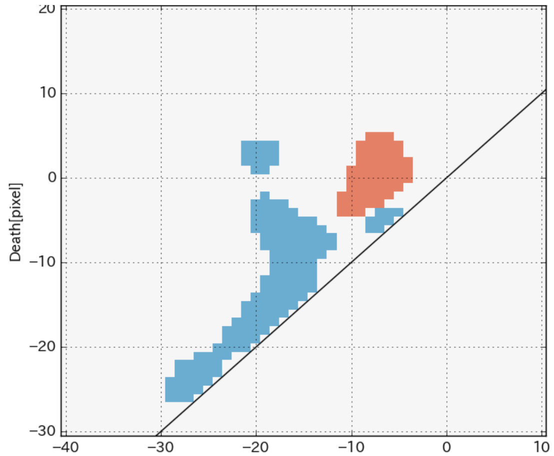
Classification result (mean accuracy) = 100%

Performance of RIDGE logistic regressions: Easy example

with Obayashi (AIMR) arXiv:1706.10082

Learned persistence diagram and its thresholding (with RIDGE)





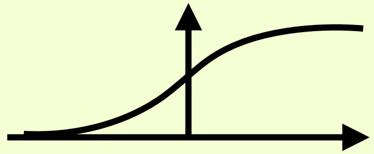
Red (resp. blue) generators contribute to 1 (resp. 0) for classification

Logistic regression model:

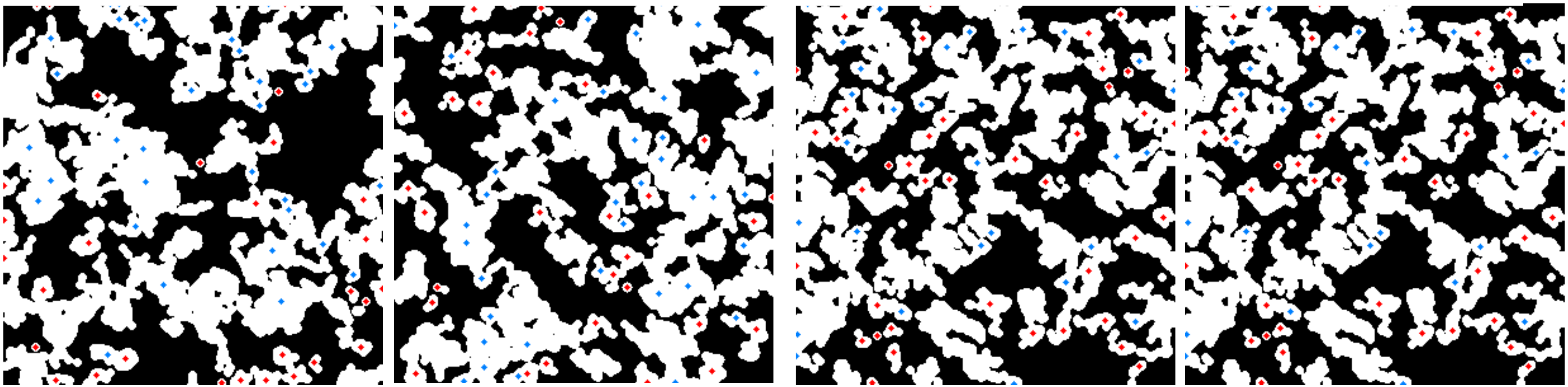
$$P(y=1 \mid w,b) = g(w\cdot x+b),$$

$$P(y=0 \mid w,b) = 1-P(y=1 \mid w,b) = g(-w\cdot x-b),$$

$$g(z)=1/(e^{-z}+1)$$



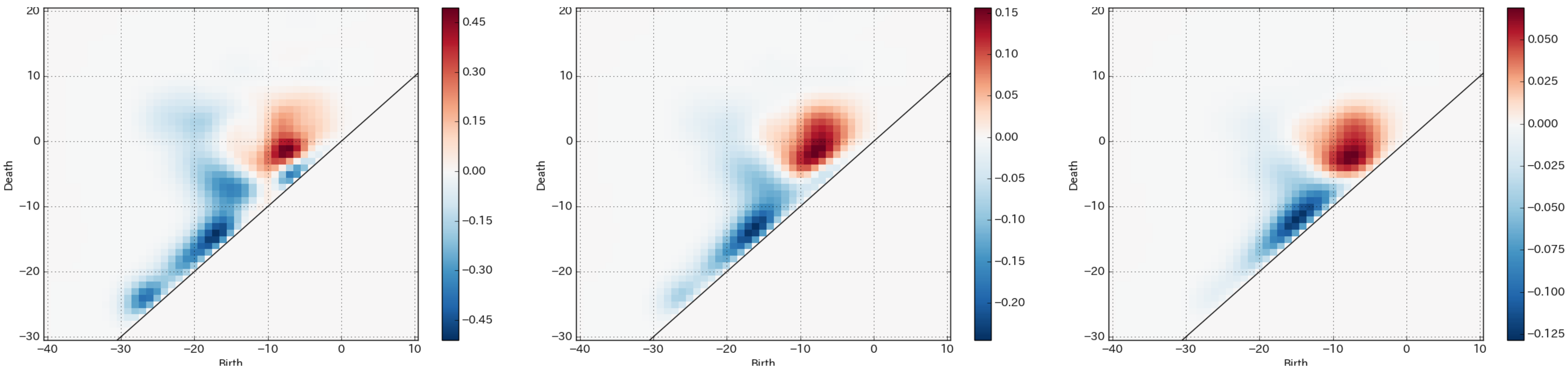
Geometric features contributing for classification (via inverse prob.)



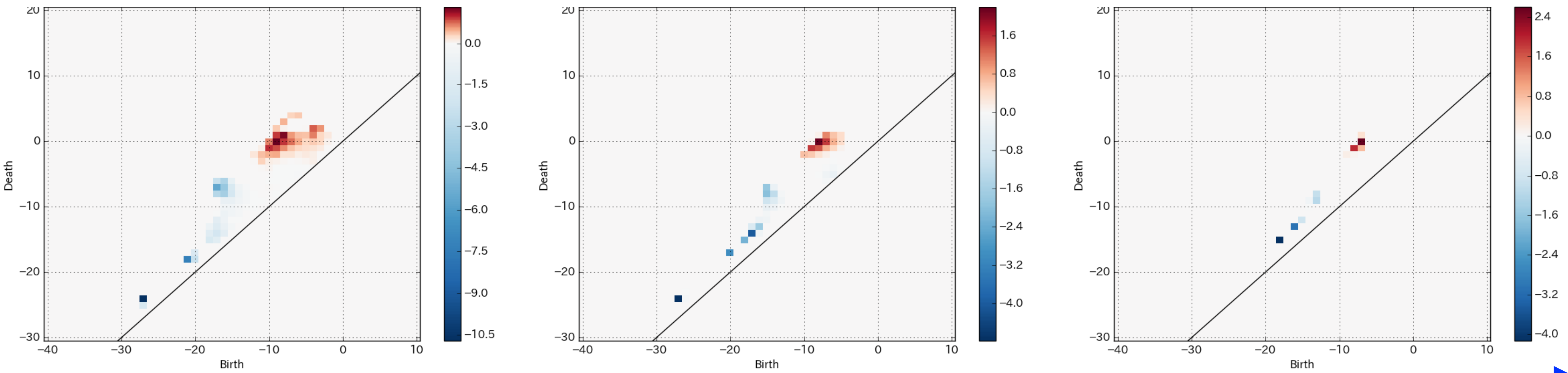
Performance of LASSO/RIDGE logistic regressions: Easy example

RIDGE/LASSO learned PDs and overfitting parameters

<RIDGE>



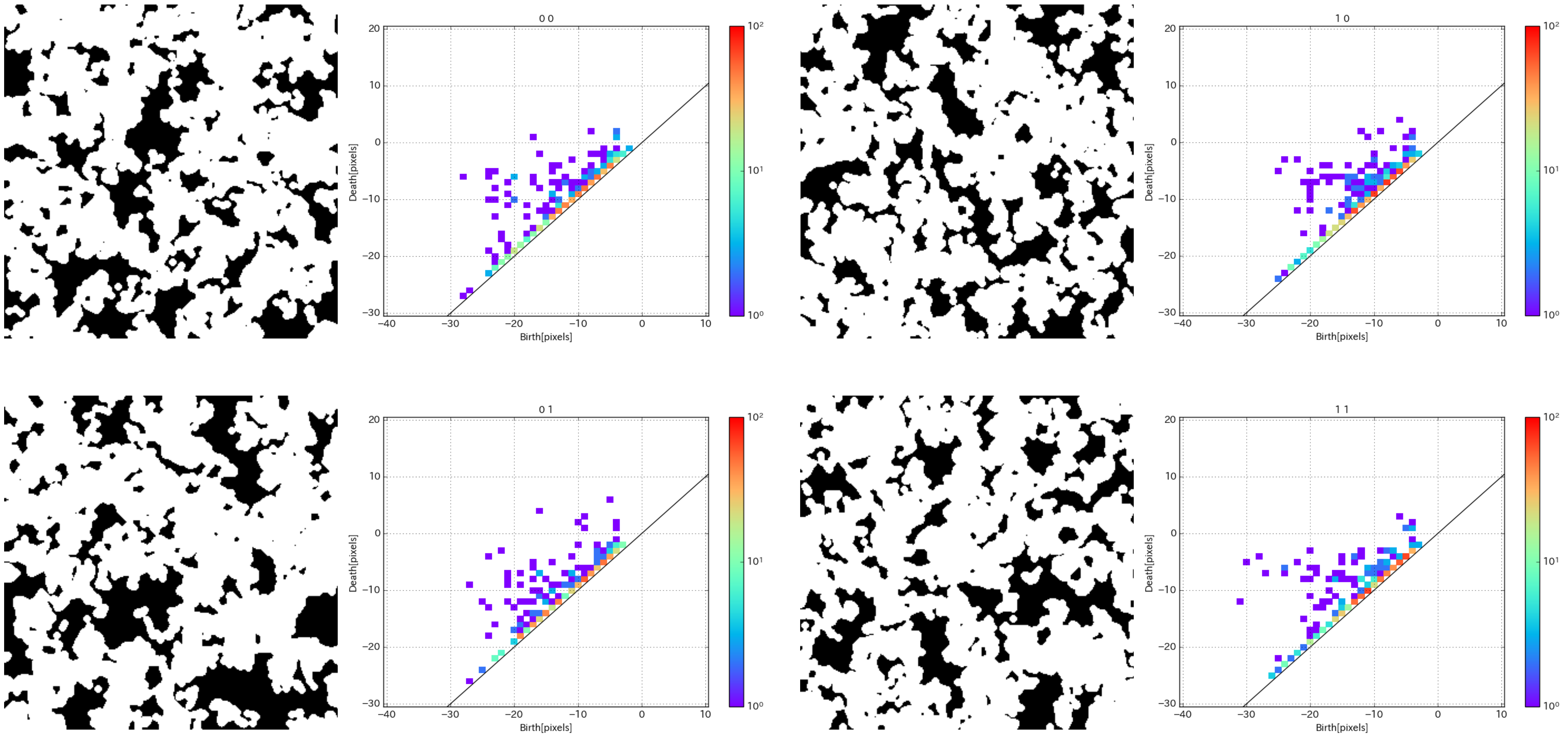
<LASSO>



(complex) (simple) λ

sparse persistence diagram shows most effective generators for learning

Performance of logistic regressions: Hard example



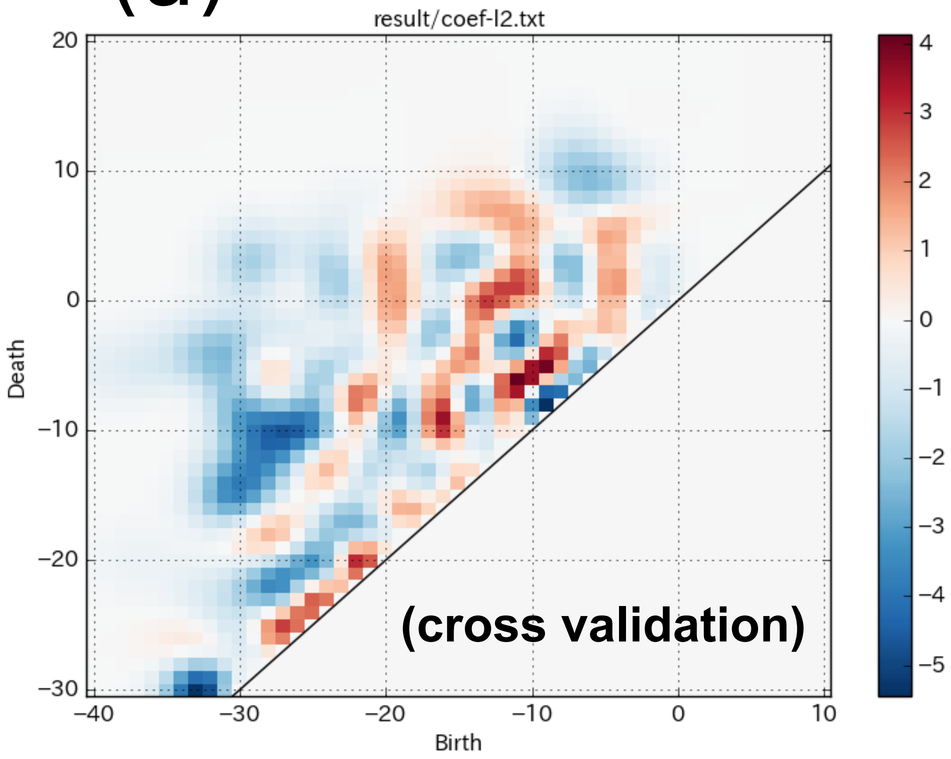
Classification result (mean accuracy) = 92%

Performance of RIDGE logistic regressions: Hard example

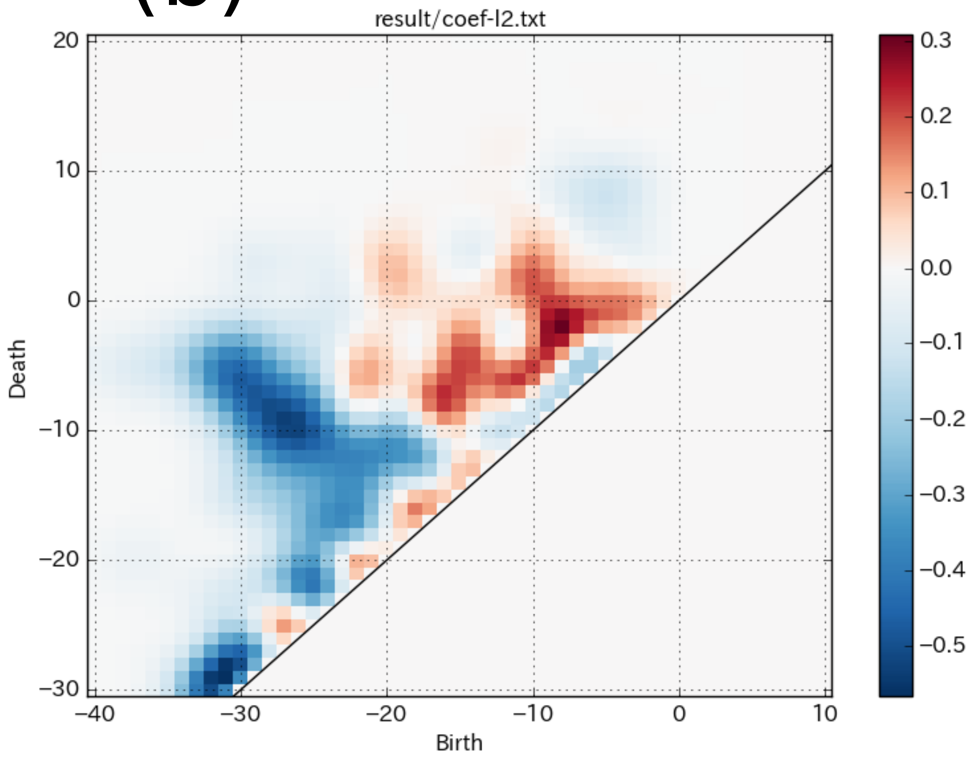
RIDGE learned PDs and overfitting parameters

<RIDGE>

(a)



(b)



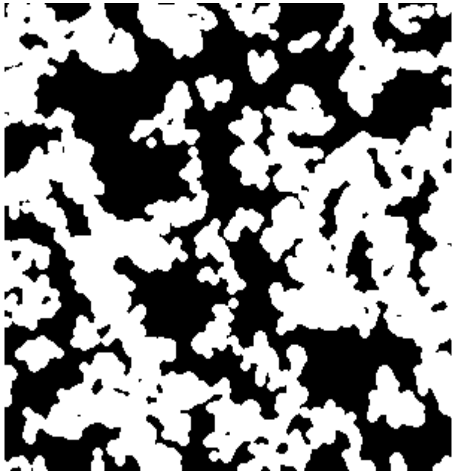
(complex) → (simple) λ

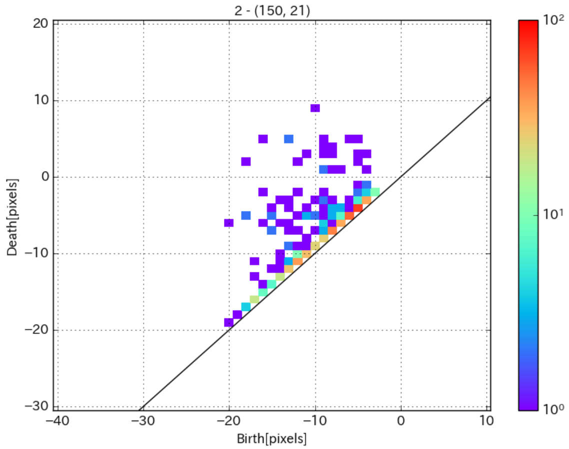
Performance comparison

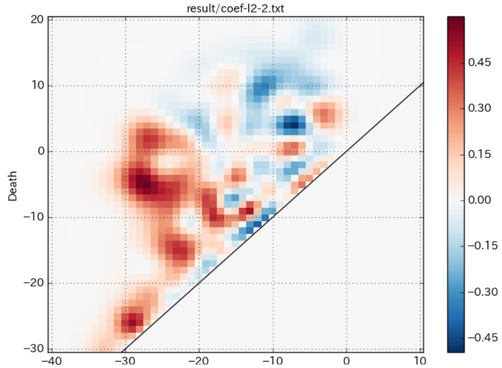
Method	Mean accuracy
PI, logistic regression, ℓ^2-penalty	0.92
PI, SVM classifier with RBF kernel	0.935
Bag of keypoints using sift with grid sampling, SVM classifier with χ^2 kernel	0.85
# of connected components of black pixels	0.73
# of connected components of white pixels	0.50
# of white pixels	0.50

Performance of linear regressions

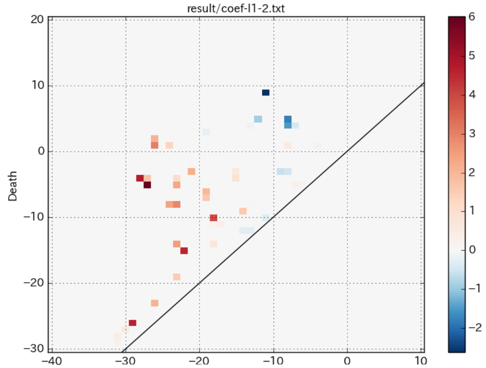
- **random images with parameters $S = 0, \dots, 9$**
- **predict S from the learned PD**



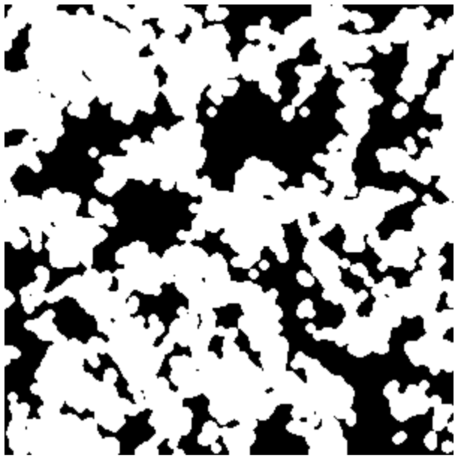


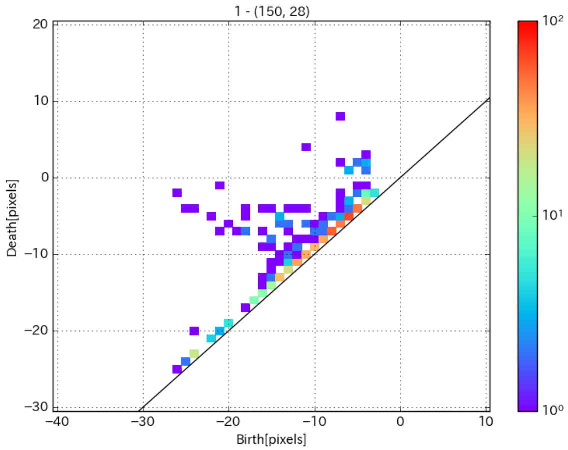


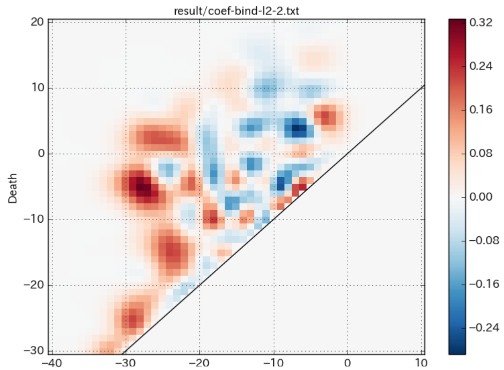
PI-RIDGE (score=0.86)



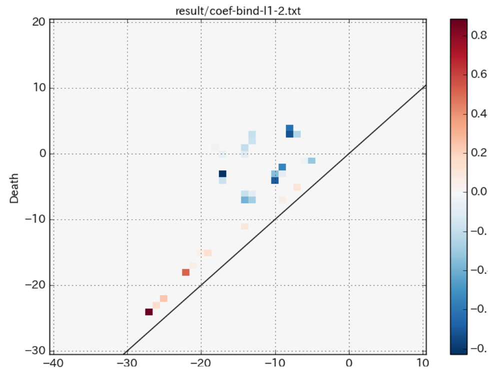
PI-LASSO (score=0.86)







Both-RIDGE (score=0.93)



Both-LASSO (score=0.94)

Conclusion

- Persistence diagrams (PD) can be a promising descriptor for materials structural analysis
- PD accepts standard inputs in materials science (point cloud and digital images)
- The software HomCloud enables an easy access to PD
- Combination of PD and ML provides a new and powerful tool for materials informatics

THANK YOU