MSRI
Mathematical Sciences Research Institute

17 Gauss Way Berkeley, CA 94720-5070 p: 510.642.0143 f: 510.642.8609 www.msri.org

# NOTETAKER CHECKLIST FORM

**(Complete one for each talk.)**

**Name:** Malgorzata Marciniak **Email/Phone:** mmarciniak@lagcc.cuny.edu 5734620411

**Speaker's Name:** Kathryn Hess

**Talk Title:** High-throughput topological screening of nanoporous materials

**Date:** 10 / 03 /2018 **Time:** 9 :30 [✓ am] / **pm (circle one)**

**Please summarize the lecture in 5 or fewer sentences:**
The database of millions of different classes of nanoporous materials, in particular zeolites, requires computational approach to tackle high-throughput screening. The goal is to find the best nano-porous materials for a given application, using a topological data analysis-based descriptor (TD) recognizing pore shapes. When some top-performing zeolites are known, TD can be used to efficiently detect other high-performing materials with high probability.

# CHECK LIST

(This is **NOT** optional, we will **not pay** for **incomplete** forms)

☑ Introduce yourself to the speaker prior to the talk. Tell them that you will be the note taker, and that you will need to make copies of their notes and materials, if any.

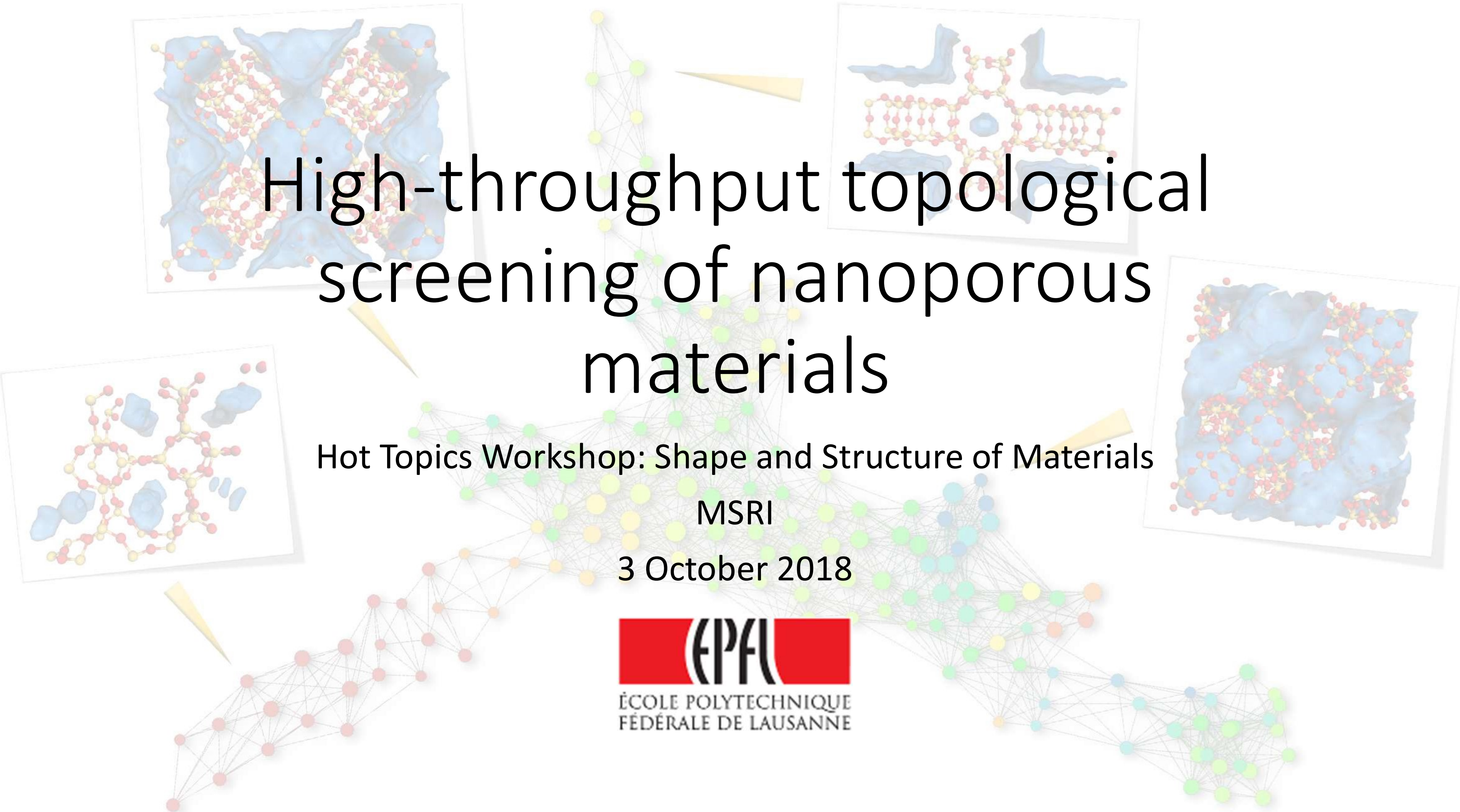
☑ Obtain ALL presentation materials from speaker. This can be done before the talk is to begin or after the talk; please make arrangements with the speaker as to when you can do this. You may scan and send materials as a .pdf to yourself using the scanner on the 3rd floor.

- **Computer Presentations**: Obtain a copy of their presentation
- **Overhead**: Obtain a copy or use the originals and scan them
- **Blackboard**: Take blackboard notes in black or blue **PEN**. We will **NOT** accept notes in pencil or in colored ink other than black or blue.
- **Handouts**: Obtain copies of and scan all handouts

☑ For each talk, all materials must be saved in a single .pdf and named according to the naming convention on the "Materials Received" check list. To do this, compile all materials for a specific talk into one stack with this completed sheet on top and insert face up into the tray on the top of the scanner. Proceed to scan and email the file to yourself. Do this for the materials from each talk.

☑ When you have emailed all files to yourself, please save and re-name each file according to the naming convention listed below the talk title on the "Materials Received" check list. (*YYYY.MM.DD.TIME.SpeakerLastName*)

☑ Email the re-named files to notes@msri.org with the workshop name and your name in the subject line.



# High-throughput topological screening of nanoporous materials

Hot Topics Workshop: Shape and Structure of Materials

MSRI

3 October 2018

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

# Collaborators

- Senja Barthel, Yongjin Lee, Seyed Mohamad Moosavi, Berend Smit (EPFL)
- Paweł Dłotko (Swansea)

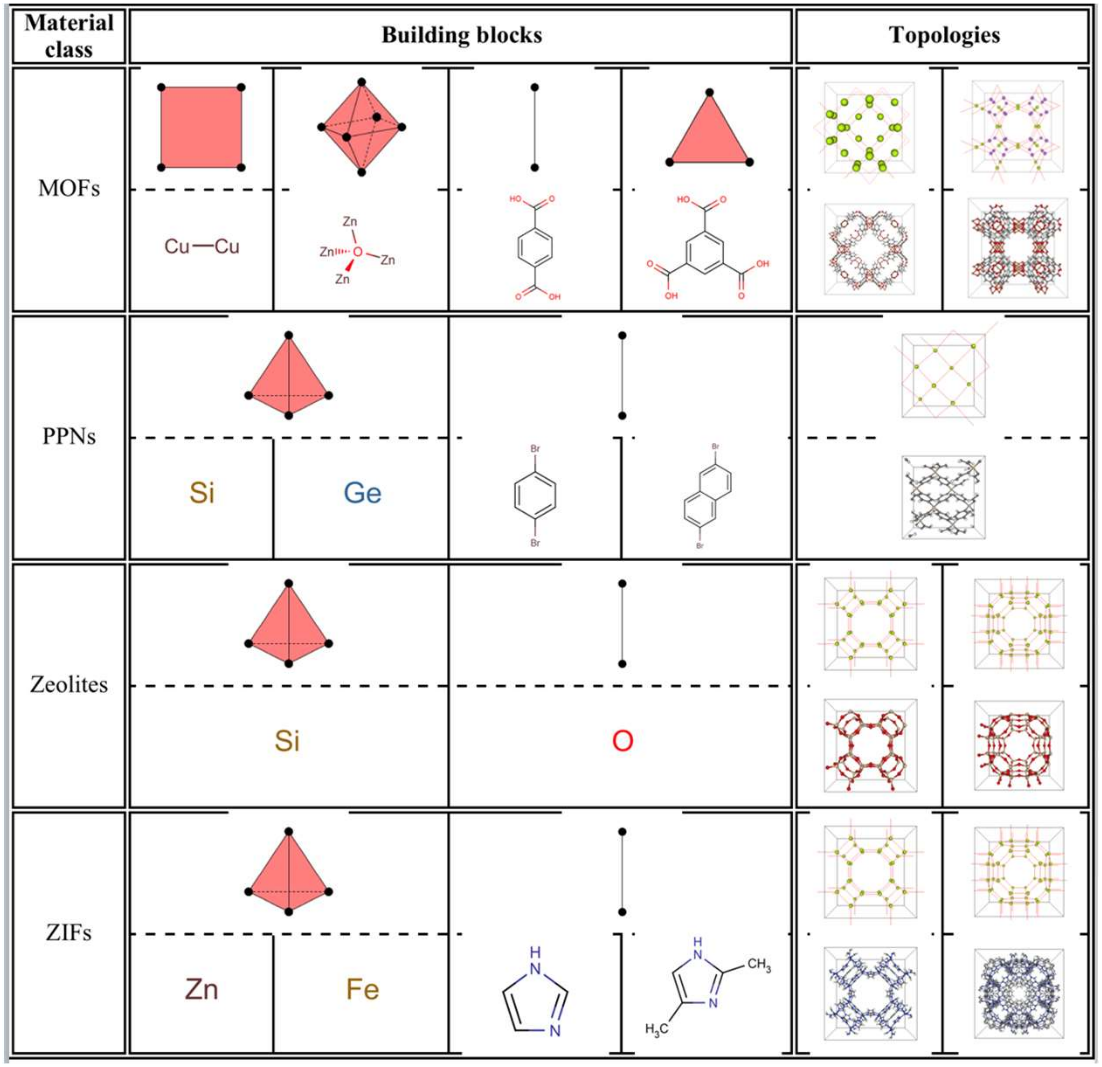
**Our articles:**

Lee, Y. et al. *Quantifying similarity of pore-geometry in nanoporous materials*. Nat. Commun. 8, 15396 doi: 10.1038/ncomms15396 (2017).

Lee, Y. et al. *High-Throughput Screening Approach for Nanoporous Materials Genome Using Topological Data Analysis: Application to Zeolites*. J. Chem. Theory Comput. 2018, 14, 4427−4437 DOI: 10.1021/acs.jctc.8b00253

# Nanoporous crystalline materials

- Over 3 million predicted structures
- Our focus
  - Zeolites:
    - 180 known
    - 300 K predicted
  - Metal organic frameworks (MOFs)
    - More than 10K synthesized
    - Over 140K predicted
- Many important applications, ranging from gas storage and separation to catalysis, sensing, etc.
- How to screen the database of possible materials to find one that is optimal for a given application, without using molecular simulation?

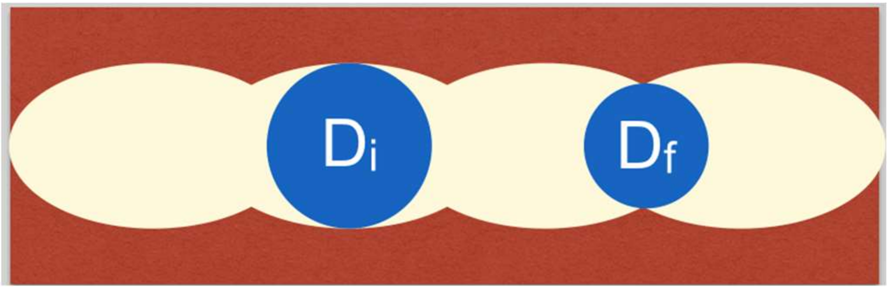


# Classical signatures

The classical signature of a nanoporous material is

$$(D_i\,, D_f\,, \rho, \mathrm{ASA}, \mathrm{AV}) \in \mathbb{R}^5.$$

- ○ $D_i$ = diameter of maximal included sphere
- ○ $D_f$ = diameter of maximal free sphere
- ○ $\rho$ = density
- ○ ASA = accessible surface area
- ○ AV = accessible volume

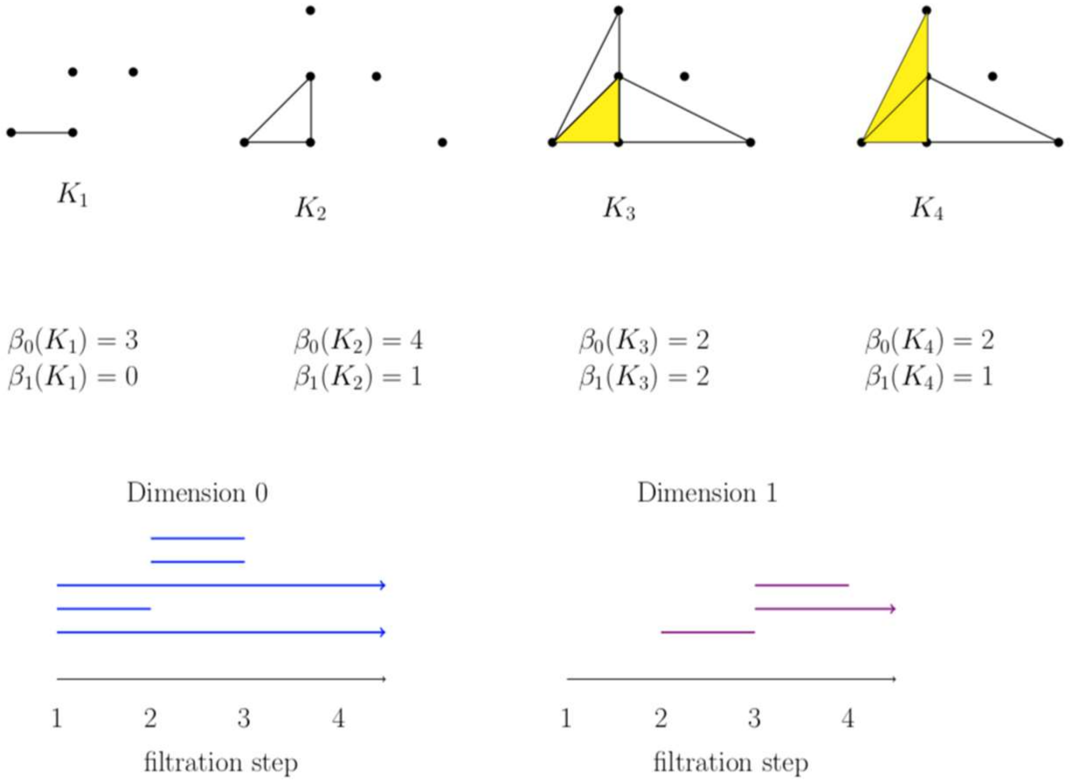


# Topological signature: overview

- Preprocessing
  - Normalization: create a supercell of each material by expanding each unit cell to approximately the size of the largest unit cell of all considered materials.
  - Extract from the software package **Zeo++** the pore system accessible to the gas molecule of interest.
  - Sample each pore system with a fixed number of points per unit surface area.

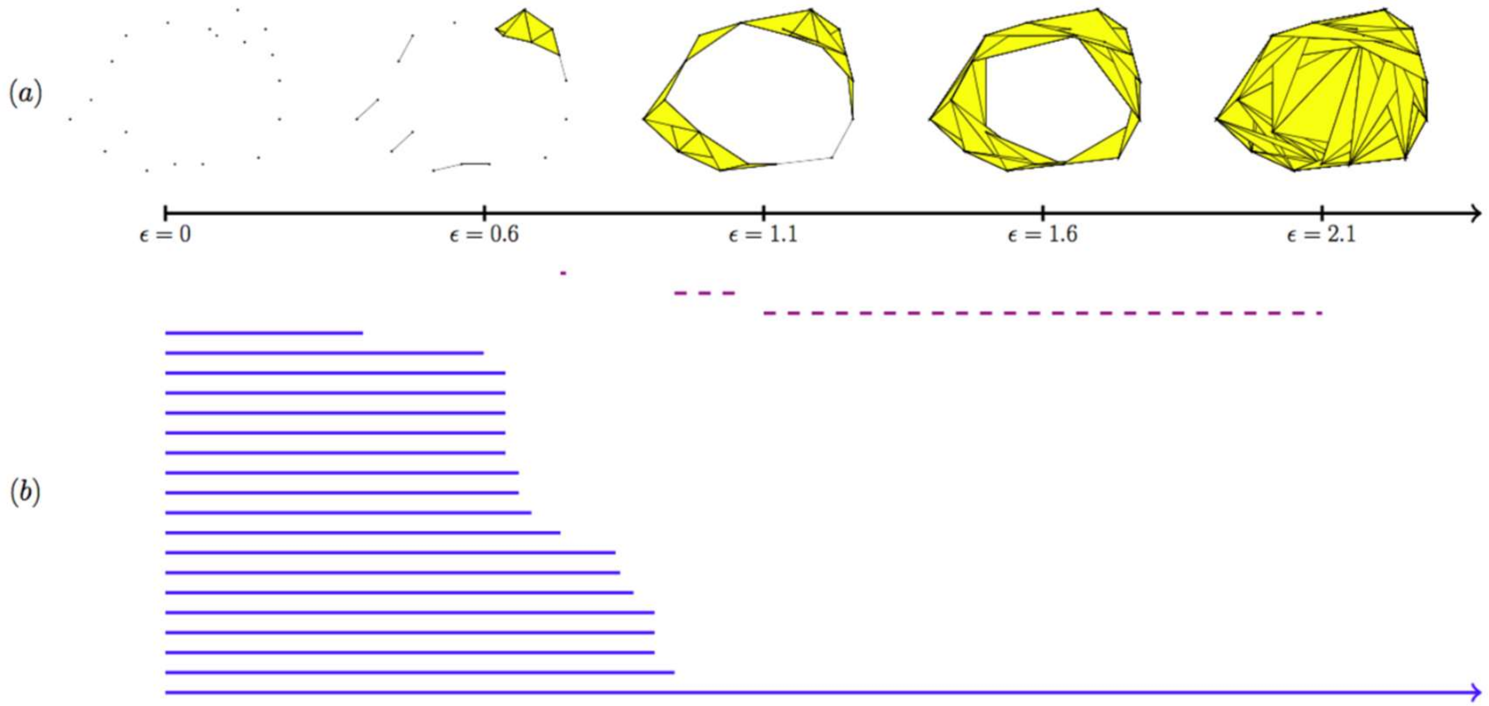
- Creation of the signature
  - Create Vietoris-Rips complexes from the sampled points, using Euclidean distance between the points.
  - Compute persistence barcodes in dimensions 0, 1, and 2.

# From point clouds to barcodes



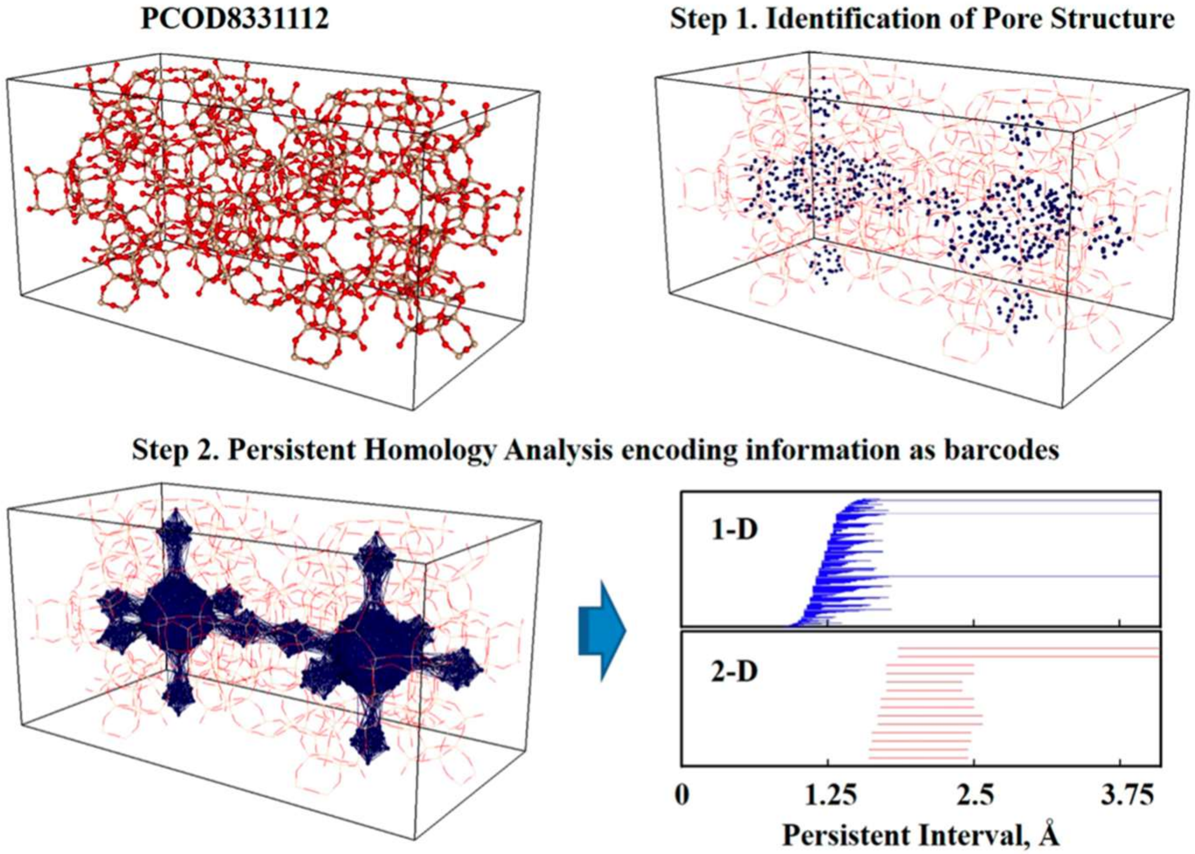
Otter et al., arXiv, 2016.

# From point clouds to barcodes



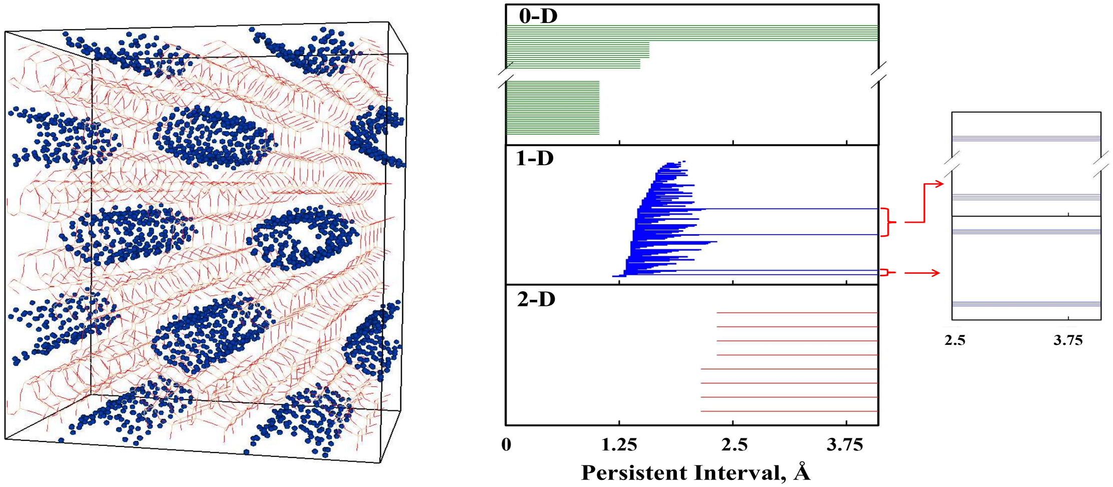
Otter et al., arXiv, 2016.

# From pore structure to signature

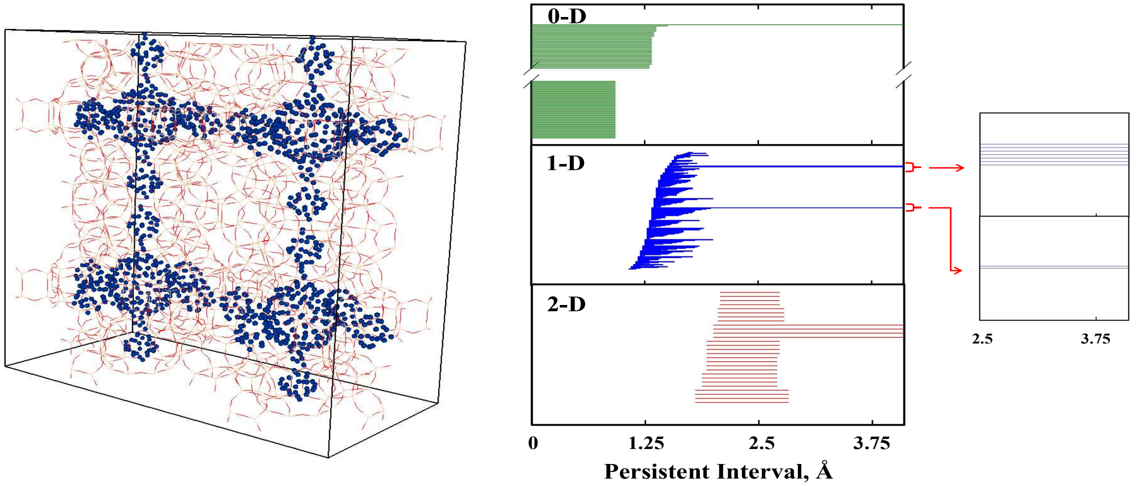

PCOD8331112
Step 1. Identification of Pore Structure
Step 2. Persistent Homology Analysis encoding information as barcodes
1-D
2-D
0
1.25
2.5
3.75
Persistent Interval, Å

# Examples of topological signatures

DON



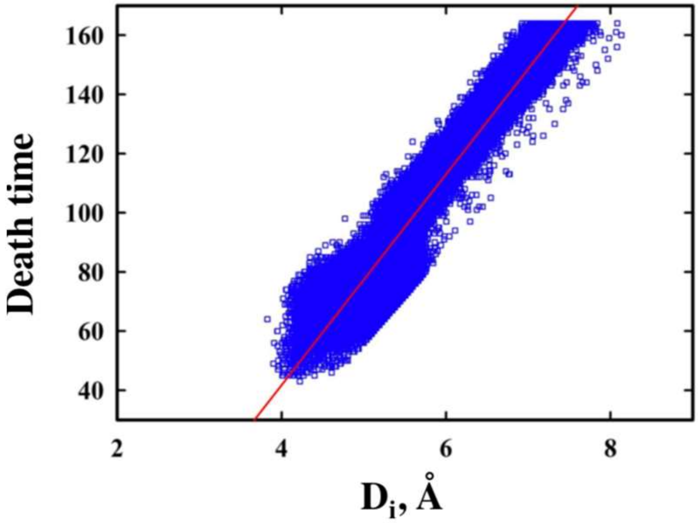
PCOD8331112



# Topological signature: details

- Zeo ++
  - detects the accessible void space inside a porous material using a periodic Voronoi network, modelling the framework atoms as hard spheres;
  - encodes the pore structure as hundreds of thousands of points on the boundary of the space where a probe molecule could be placed.
- To sample points provided by Zeo++
  - combine random sampling and grid sampling;
  - for the random sampling, choose one point per 2 Å$^2$ surface area, at least 0.8Å from all other sampled points;
  - for each grid cube (side length 0.5 Å), choose the point that is closest to the midpoint of the cube and add it to the random sampling if its distance to the randomly sampled points is greater than 0.8 Å.

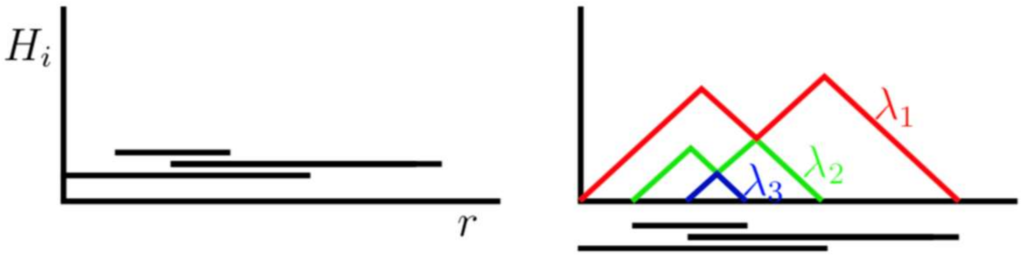
# Topological signature: details



- The Vietoris-Rips complex
  - grown in 164 steps of 0.025 Å increments, from 0 to 4.1 Å;
  - bound prevents geodesically distant points of the surface that are close in Euclidean metric from being connected;
  - describes the *embedding* of the pore surface into the ambient space;
  - not all homology classes die by maximal filtration: assign a maximum death time, based on linear fit between $D_i$ and the death time for smaller pores.

# Persistence landscapes

- Barcodes give rise to *persistence landscapes*.



$$\lambda = \left\{ \lambda_k : \mathbb{R} \to \mathbb{R} \cup \{\infty\} \mid k \in \mathbb{N} \right\}$$

- The *L2-landscape distance* between barcodes B and B’ with associated landscapes **λ** and **λ**’:

$$\Lambda(B, B') = \|\lambda - \lambda'\|_2 = \sum_{k=1}^{\infty} \left( \int |\lambda_k(t) - \lambda'_k(t)|^2 dt \right)^{\frac{1}{2}}$$

Bubenik, J Mach Learn Res (2015)
Dlotko & Bubenik, J Symbolic Comp (2017)

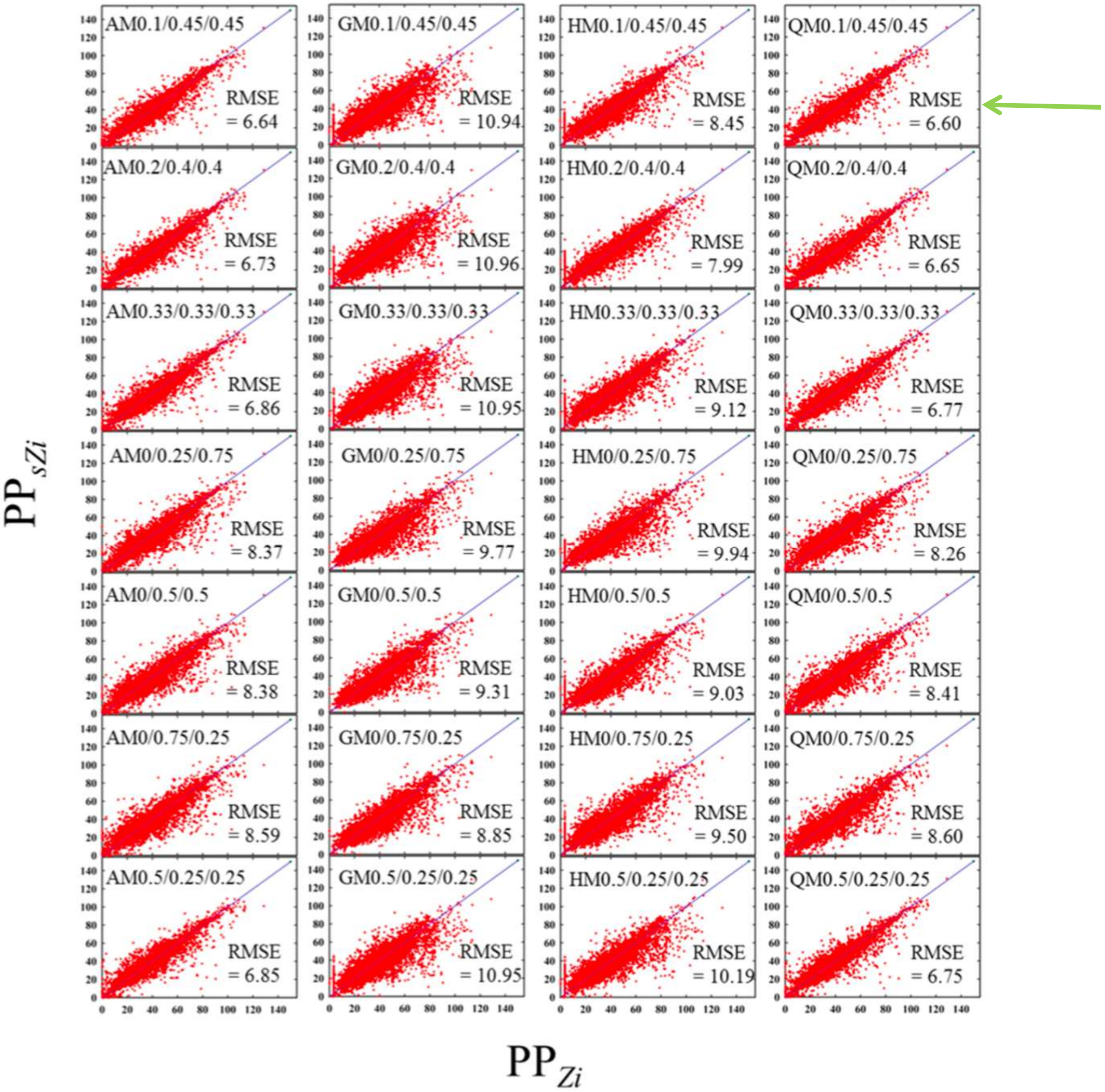
# Distance between topological signatures: $D_{TS}$

$$D_{\mathrm{TS}} := \sqrt{\alpha_0 L_0^2 + \alpha_1 \Lambda_{d=1}^2 + \alpha_2 \Lambda_{d=2}^2},$$

$$L_0 = \left|\frac{n_1}{V_1} - \frac{n_2}{V_2}\right|$$

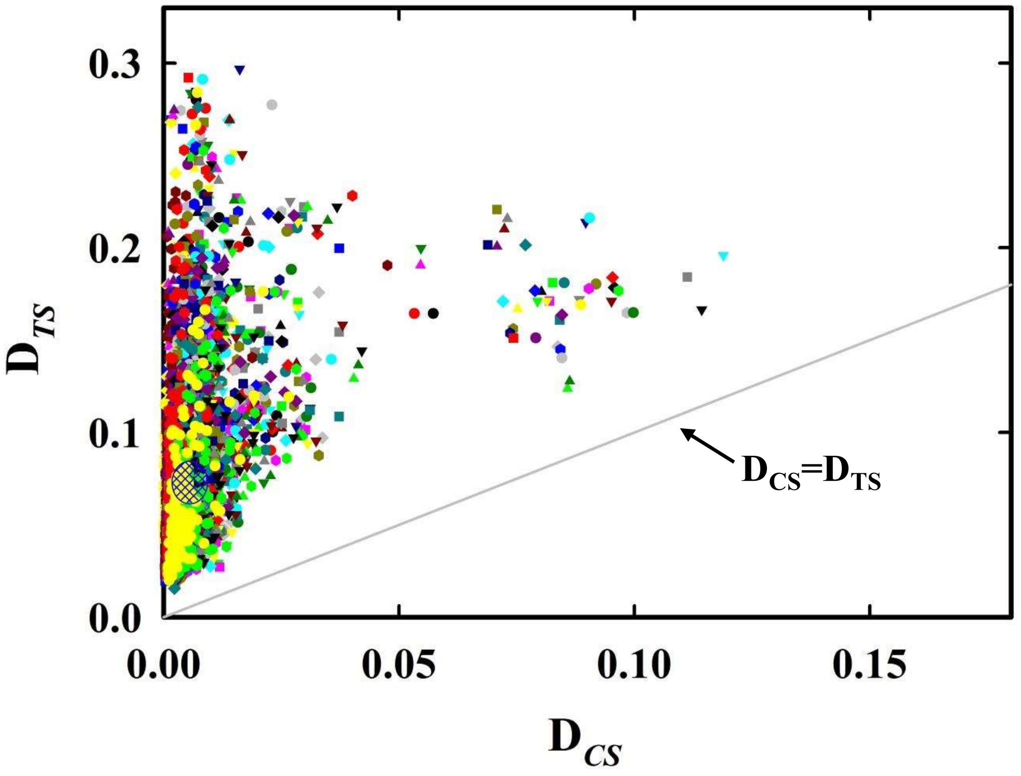
- $\boldsymbol{\alpha}_0$= 0.1, $\boldsymbol{\alpha}_1$= 0.45, and $\boldsymbol{\alpha}_2$ = 0.45: minimize the error in predicting global structural properties and performance properties for a test set of 5000 materials.
- $L_2$-distances chosen, instead of $L_p$ for some other p, for similar reasons.
- $n_i$ = the number of points sampled on material i.
- $V_i$ = the volume of its supercell.

# Choosing the weights

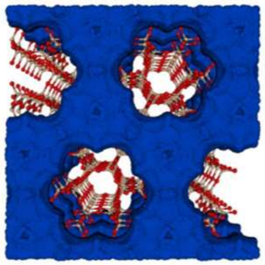
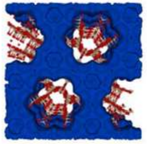
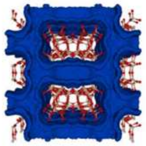
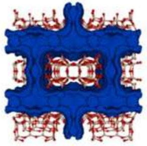
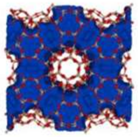
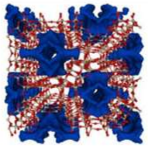
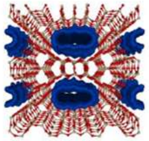
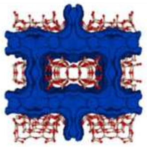
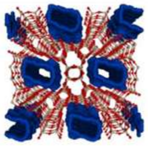
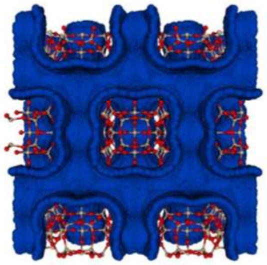
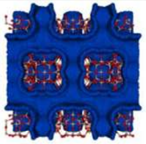
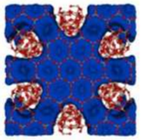
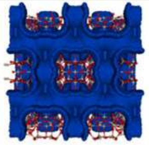
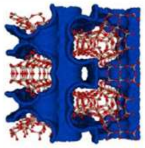
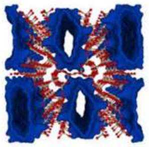
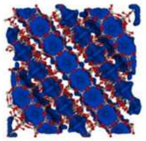


# Distance comparison

The distance $D_{CS}$ between two classical signatures is the Euclidean distance between the vectors. Modifying the weights has little to no effect on the relation between $D_{CS}$ and $D_{TS}$.

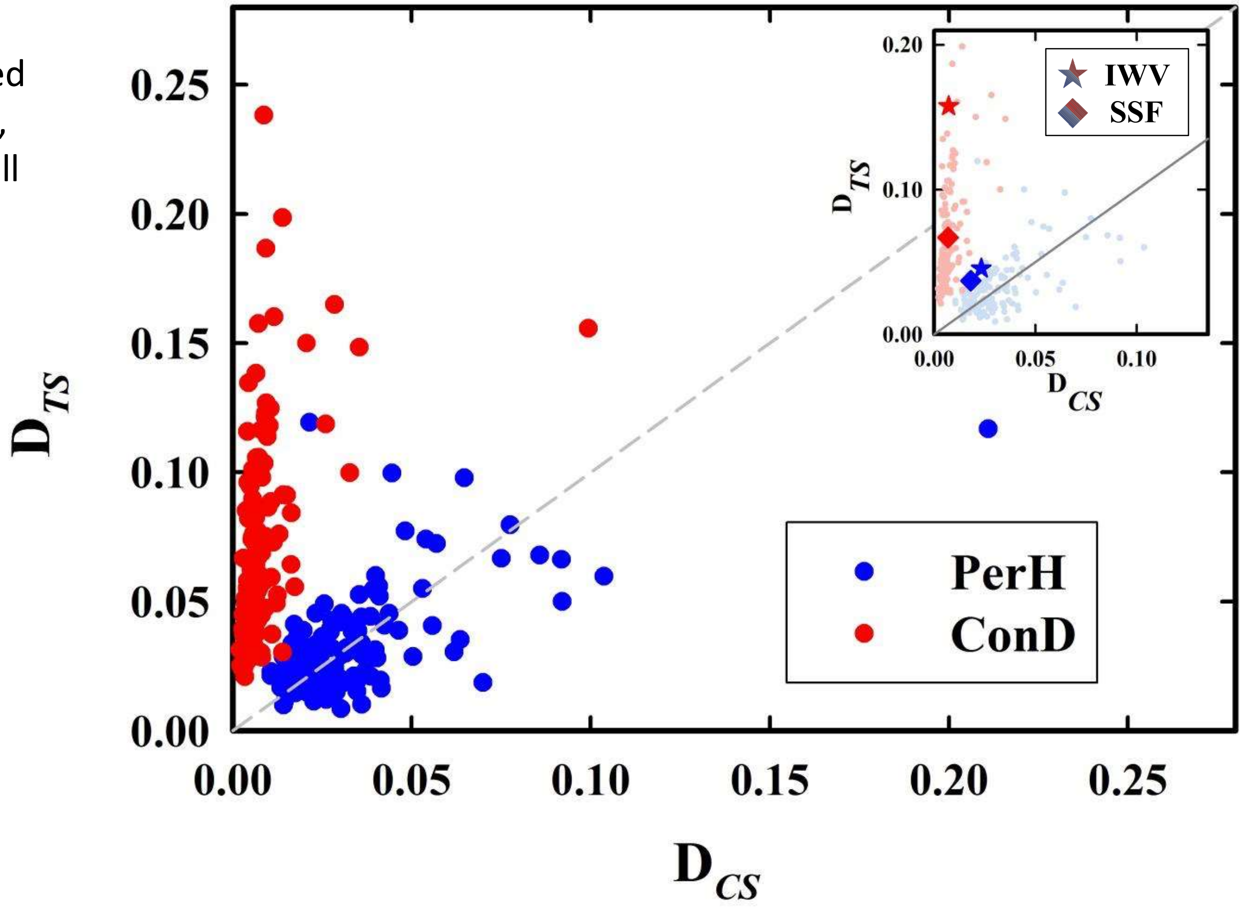


# Four most similar zeolite structures

| Seed | Descriptor | Selected Nth Similar Structure | | | |
|---|---|---|---|---|---|
| | | 1st | 2nd | 3rd | 4th |
| SSF<br> | PerH |  |  |  |  |
| | ConD |  |  |  |  |
| IWV<br> | PerH |  |  | |  |
| | ConD | |  |  |  |

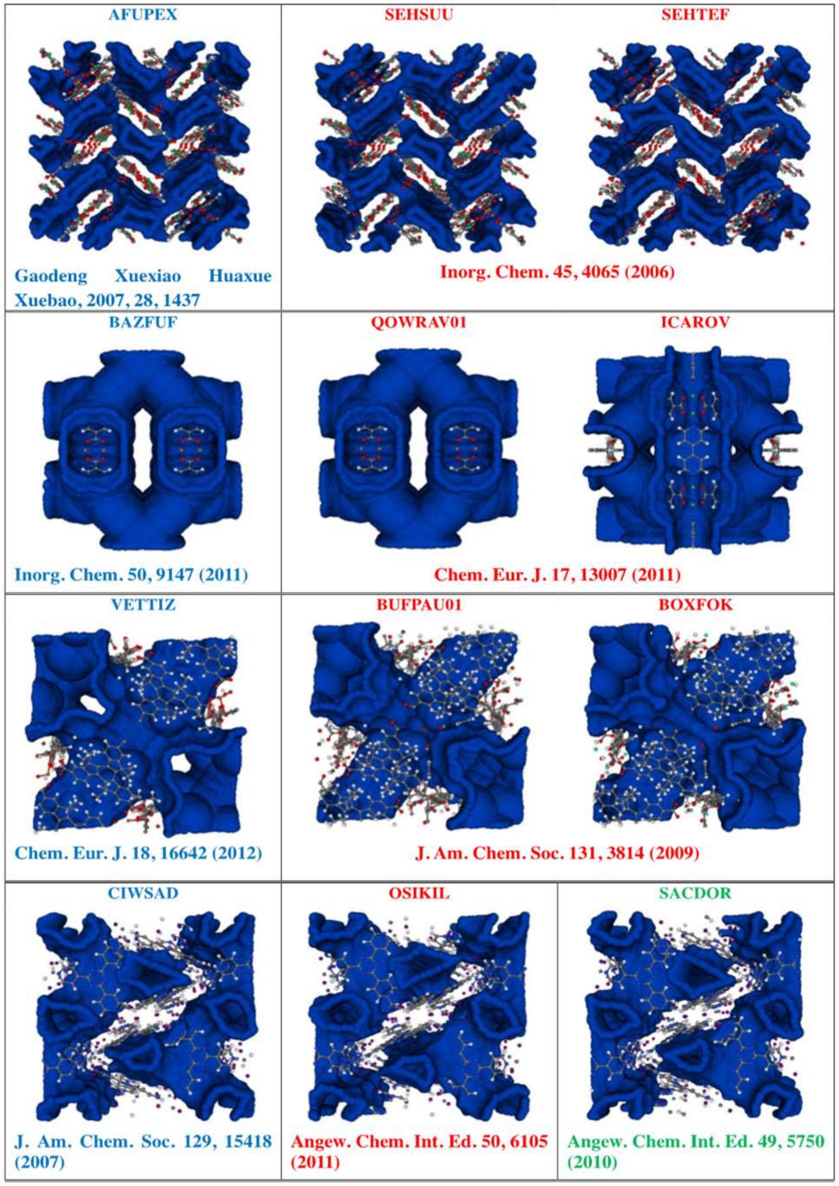
# Four most similar zeolite structures

**Conclusion:** When topology is used to identify similar pore structures, small $D_{TS}$ correlates well with small $D_{CS}$, better than when classical signature used.



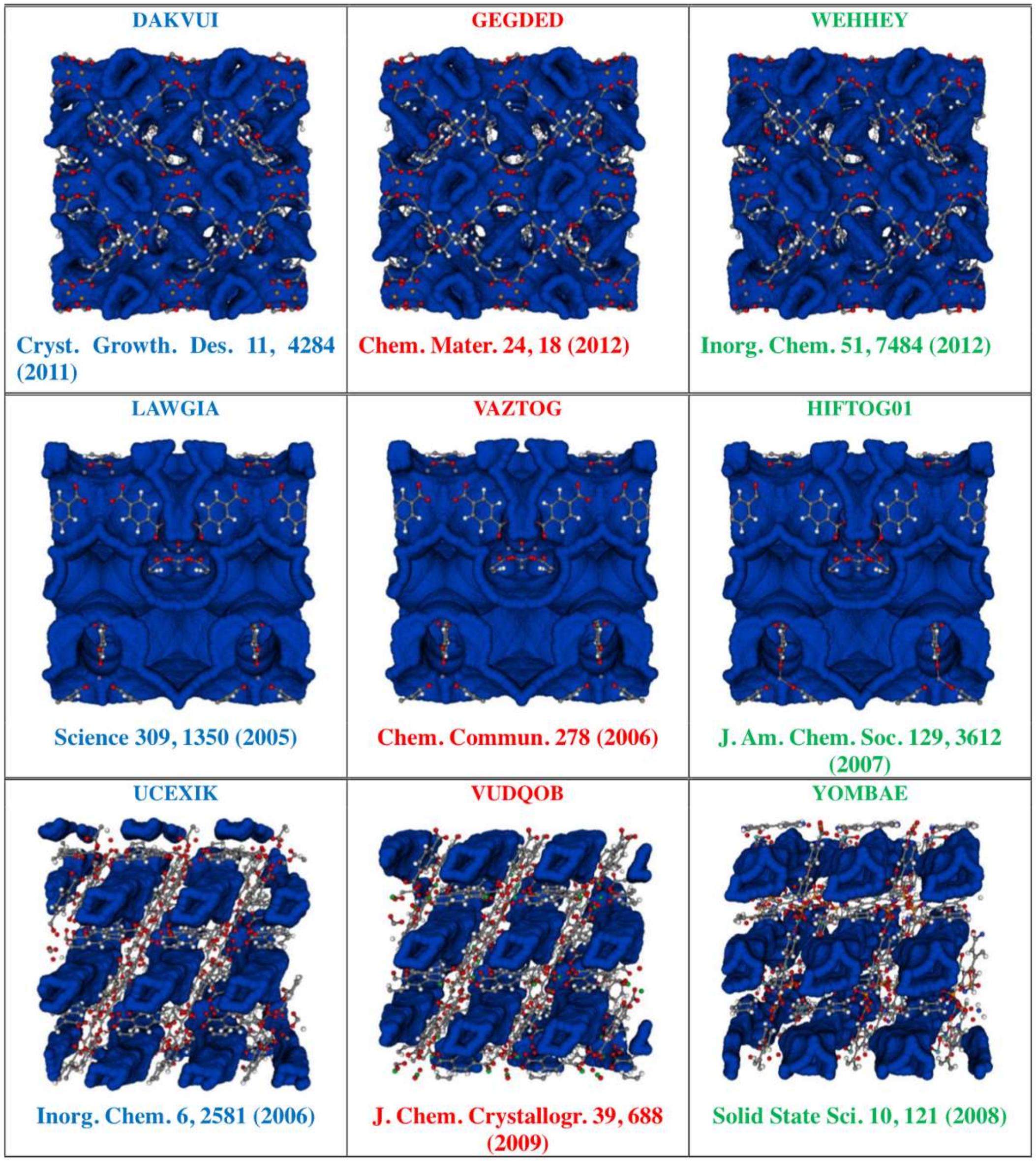
# MOFs with similar geometry

(Similarities unreported in the literature)


AFUPEX
Gaodeng Xuexiao Huaxue Xuebao, 2007, 28, 1437
SEHSUU
SEHTEF
Inorg. Chem. 45, 4065 (2006)
BAZFUF
Inorg. Chem. 50, 9147 (2011)
QOWRAV01
ICAROV
Chem. Eur. J. 17, 13007 (2011)
VETTIZ
Chem. Eur. J. 18, 16642 (2012)
BUFPAU01
BOXFOK
J. Am. Chem. Soc. 131, 3814 (2009)
CIWSAD
J. Am. Chem. Soc. 129, 15418 (2007)
OSIKIL
Angew. Chem. Int. Ed. 50, 6105 (2011)
SACDOR
Angew. Chem. Int. Ed. 49, 5750 (2010)

# MOFs with similar geometry

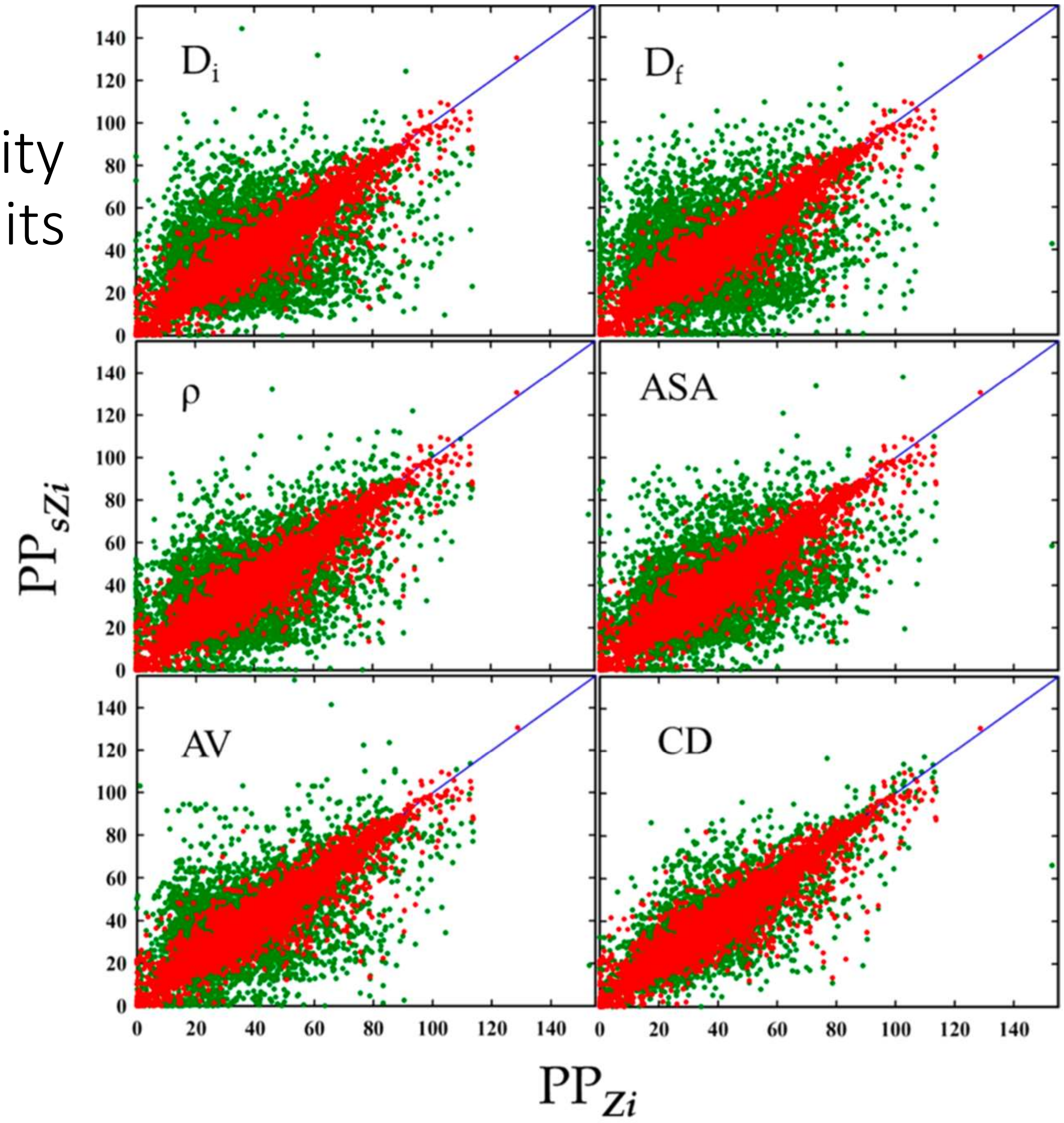
(Similarities unreported in the literature)


DAKVUI
Cryst. Growth. Des. 11, 4284 (2011)
GEGDED
Chem. Mater. 24, 18 (2012)
WEHHEY
Inorg. Chem. 51, 7484 (2012)
LAWGIA
Science 309, 1350 (2005)
VAZTOG
Chem. Commun. 278 (2006)
HIFTOG01
J. Am. Chem. Soc. 129, 3612 (2007)
UCEXIK
Inorg. Chem. 6, 2581 (2006)
VUDQOB
J. Chem. Crystallogr. 39, 688 (2009)
YOMBAE
Solid State Sci. 10, 121 (2008)

# The methane storage problem

- **Goal:** a topology-based methodology to quantify similarity of the chemical environment of adsorbed molecules, in order to develop computationally feasible high-throughput screening for high-performance materials.
- **Relevant performance property:** *deliverable capacity,* i.e., the difference in loading (number of methane molecules per unit material) at the (high) pressure at which we charge the materials with methane and at the (low) pressure at which we discharge the material.

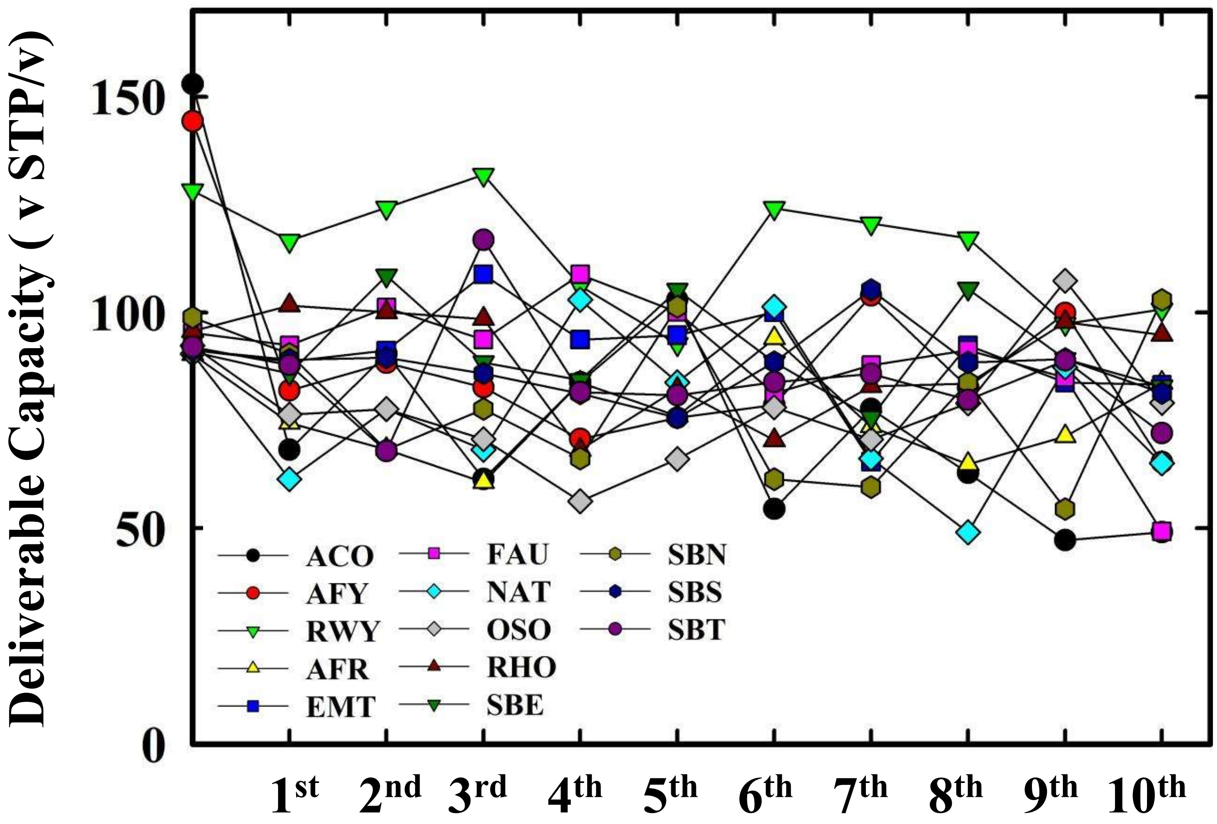
Deliverable capacity of each zeolite vs its closest match



Red = TS
Green = classical

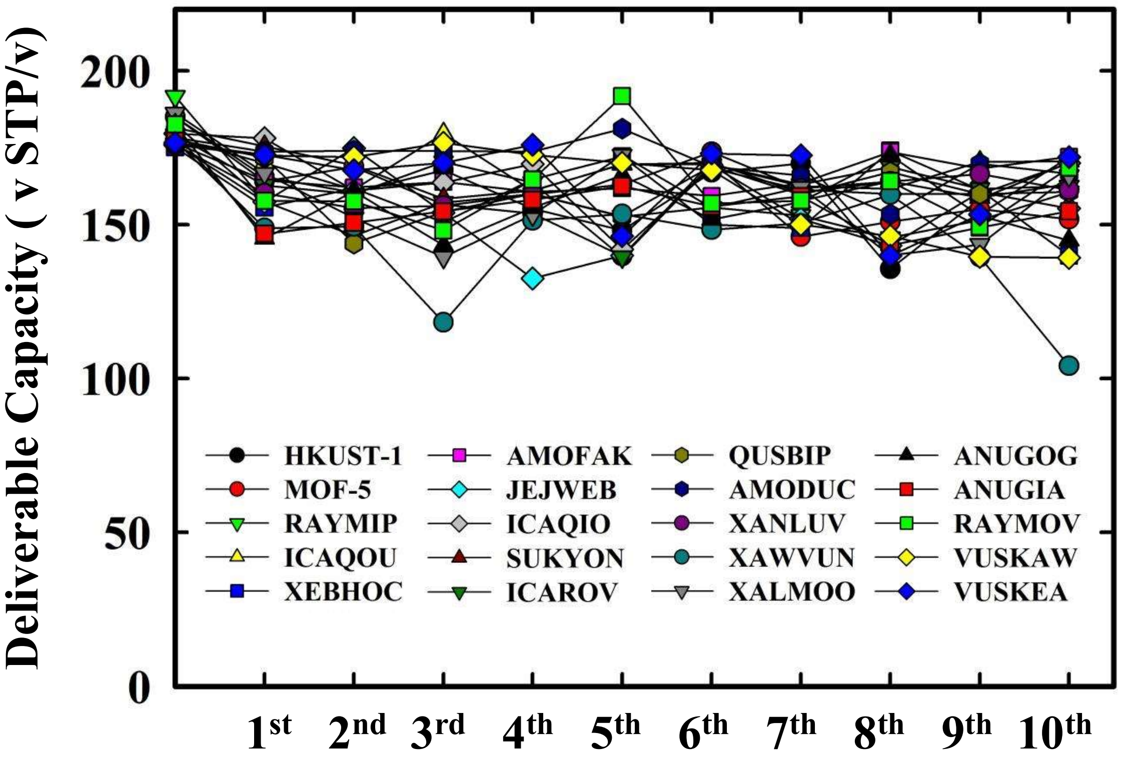
# Performance properties of structures similar to top 13 zeolites

80% have similarly high deliverable capacity



# Performance properties of structures similar to top 20 MOFs

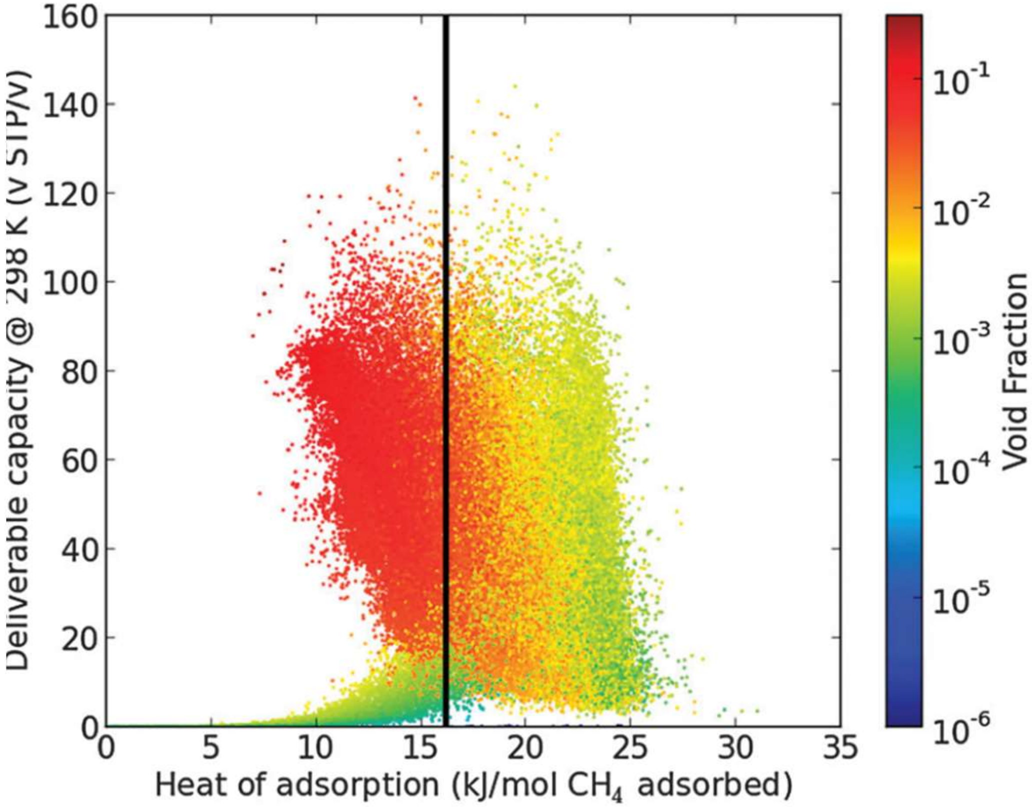
85% have similarly high deliverable capacity



# Optimal conditions for adsorptive storage?

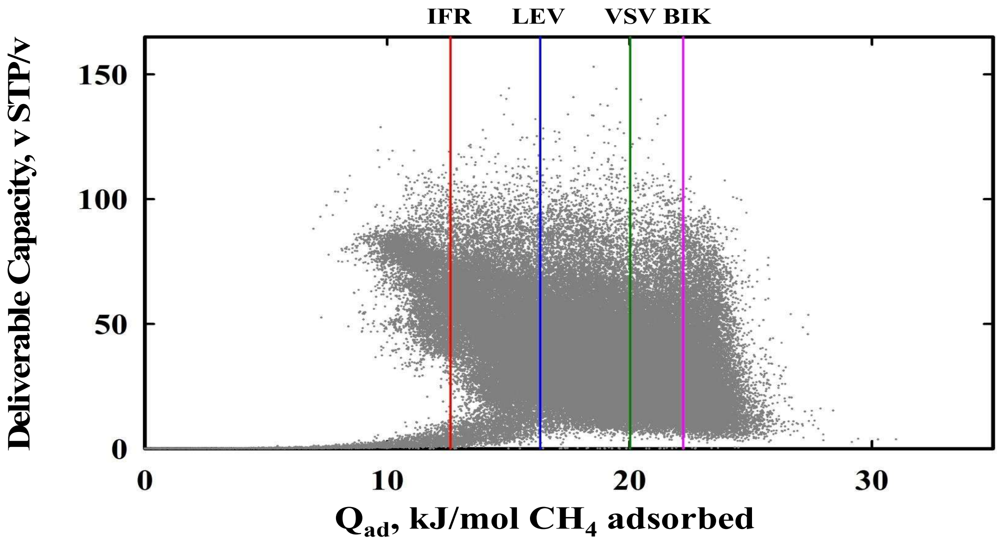
**Claim:** [Bathia-Meyers, 2006] There is an optimal heat of adsorption that maximizes deliverable capacity of a nanoporous material for methane storage.

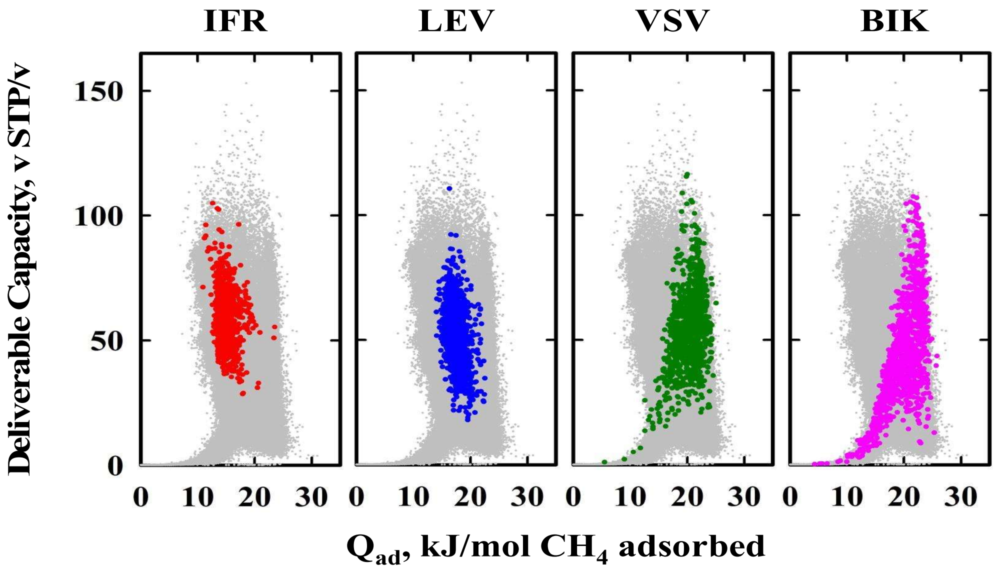
**Question:** How to relate this claim to the molecular simulation results of [Simon et al., 2014] plotted on the right?



# Deliverable capacity and heat of adsorption

- **Top:** all zeolites [Simon et al., 2014]
- **Bottom:** 500 geometrically most similar structures to four references structures.
- **Conclusion:** There is not a single class of optimal materials





# Mapper: another TDA tool

- Unsupervised mutivariate pattern analysis of high-dimensional data, retains more information than PCA
- Produces a compressed visual representation of the data, providing a strong indication of where to look for meaningful clustering and encoding relations between clusters
- Numerous remarkably successful applications, e.g., to the discovery of a new subtype of breast cancer [Levine et al., PNAS 2011].

# Mapper: another TDA tool

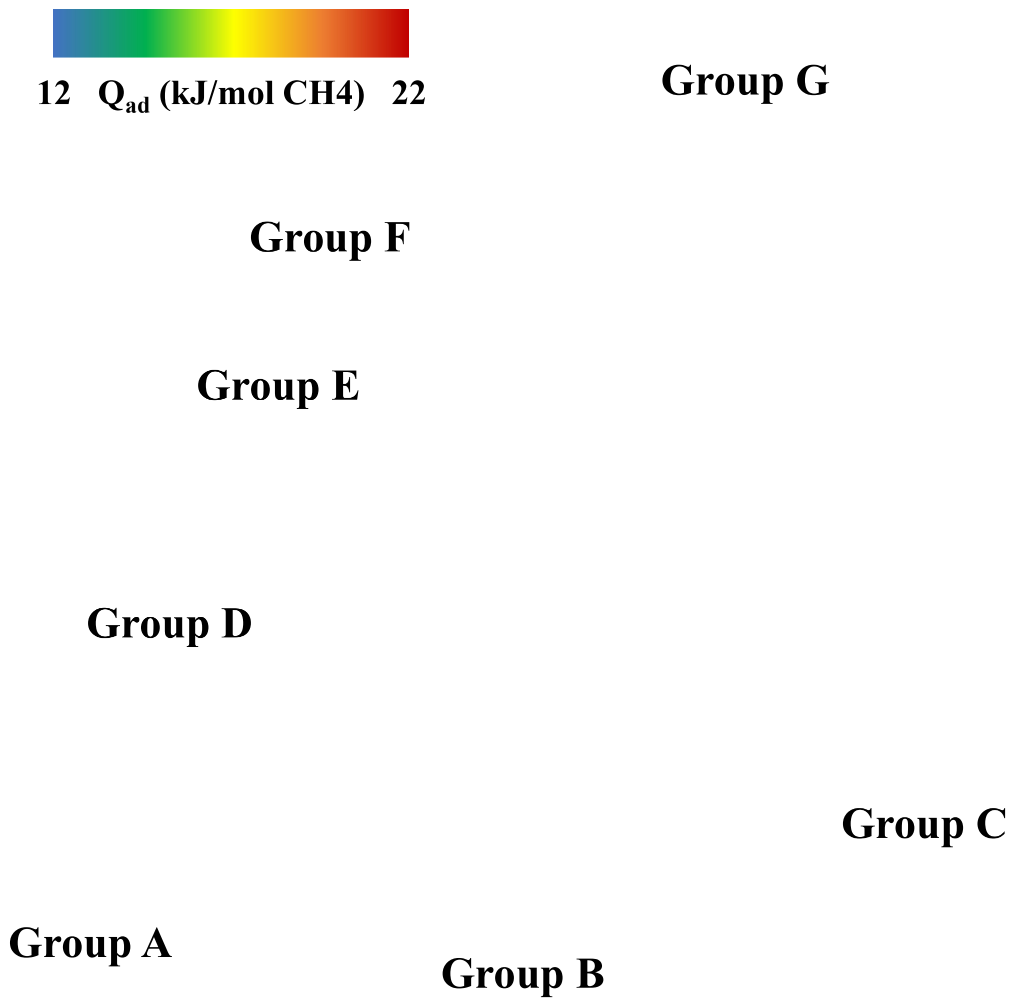
- **Input:**
  - Data set X equipped with notion of “distance” between points
  - Function f: X → $\mathbb{R}^n$
  - Cover of $\mathbb{R}^n$
- **Process:**
  - Pull the cover back to X via f.
  - Cluster points in the pre-images of the opens in the cover, usually by single-linkage clustering.
  - Visualize: clusters as nodes, connected by an edge if they share a common element and colored by some relevant average value.

# Mapper plot of top 1% of zeolites for methane storage

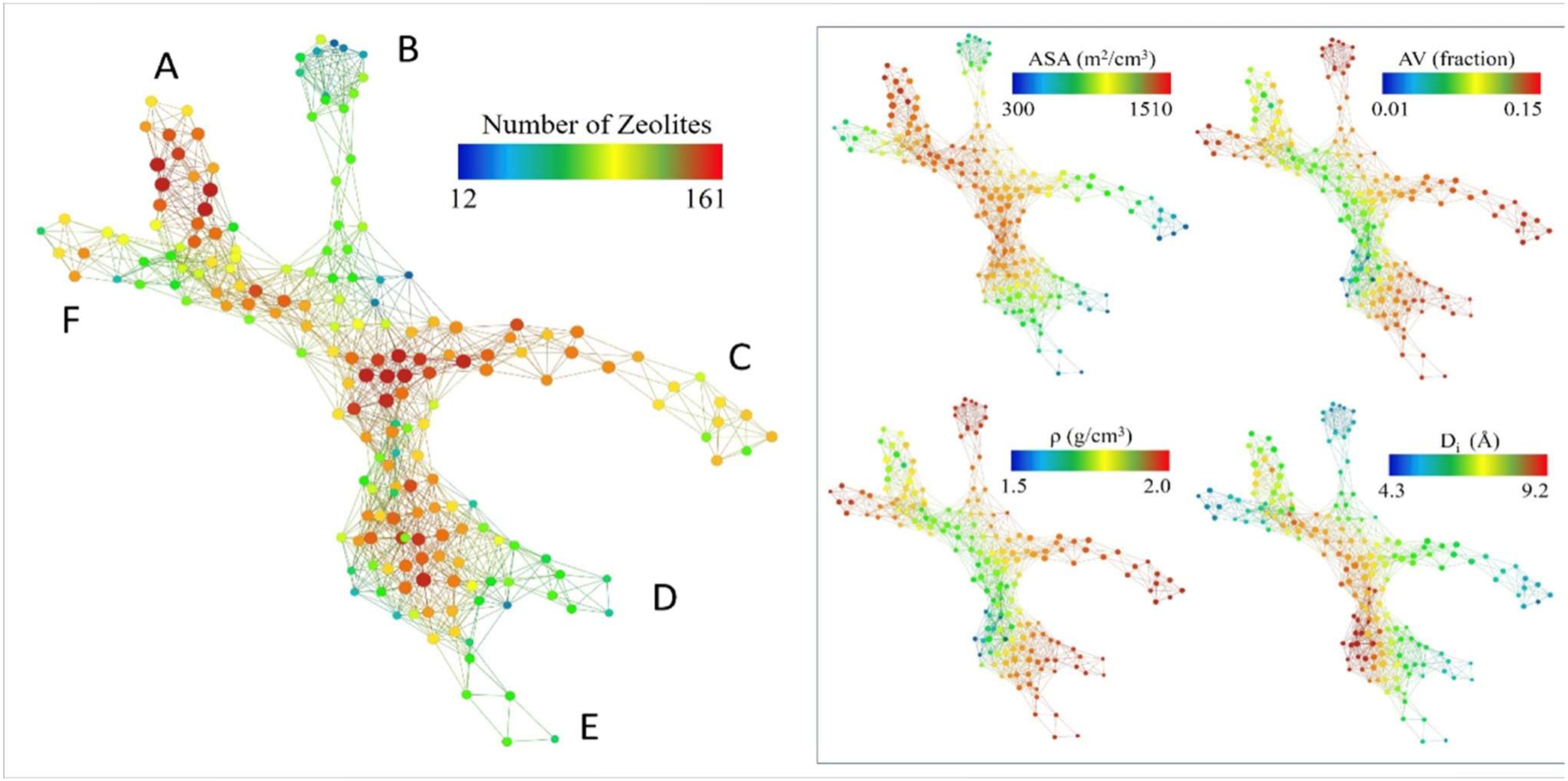
Distance = distance between topological signatures

Nodes are colored by mean value of heat of adsorption (red = high, blue = low)

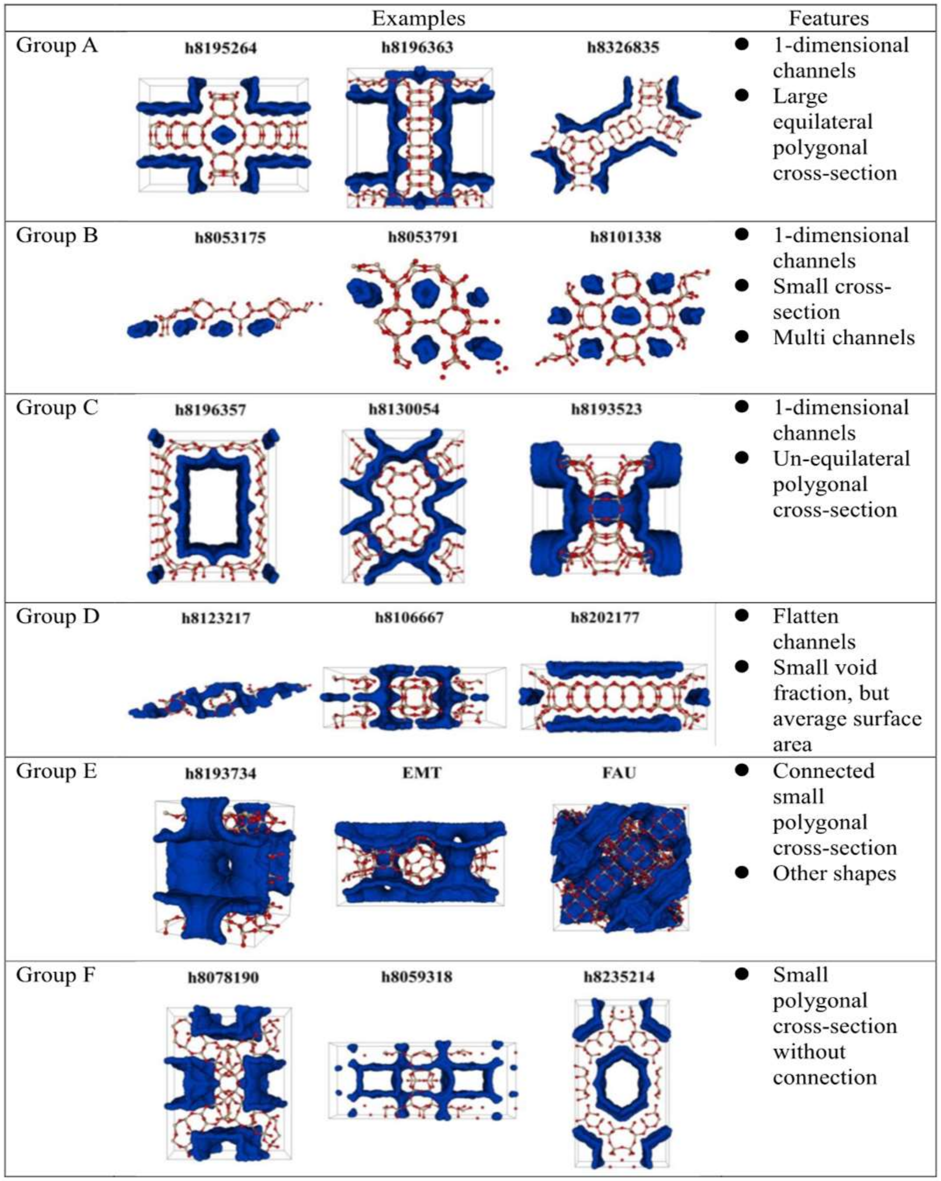
(Obtained with the Ayasdi Core software platform (www.ayasdi.com).



# Diversity of zeolites

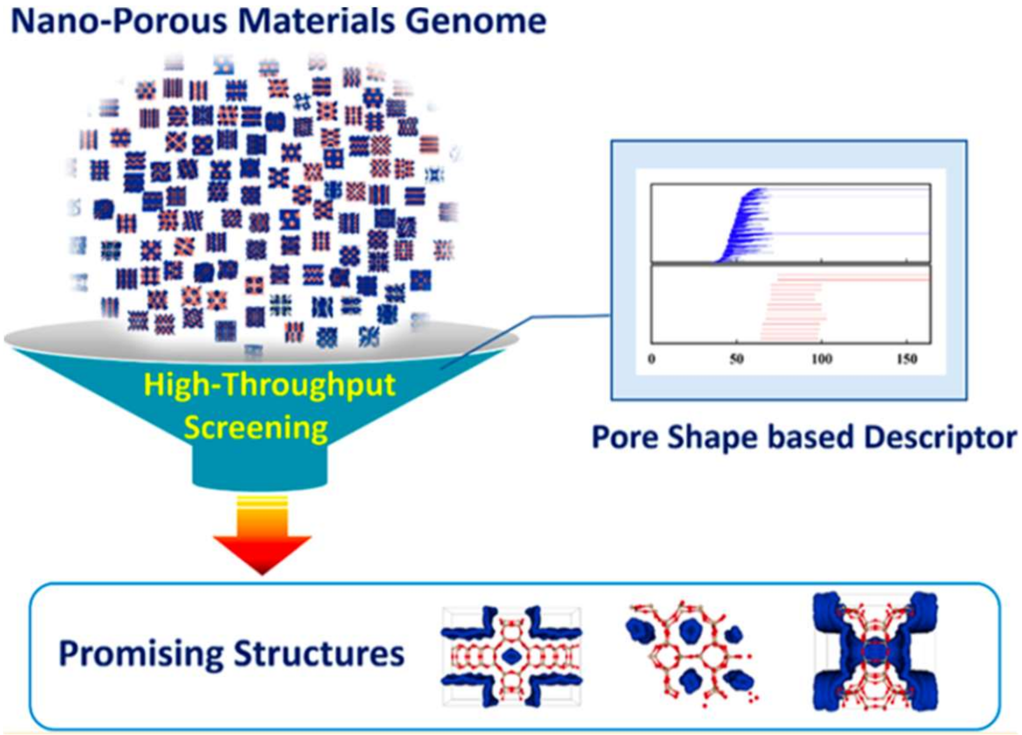


# Examples and features of the six groups

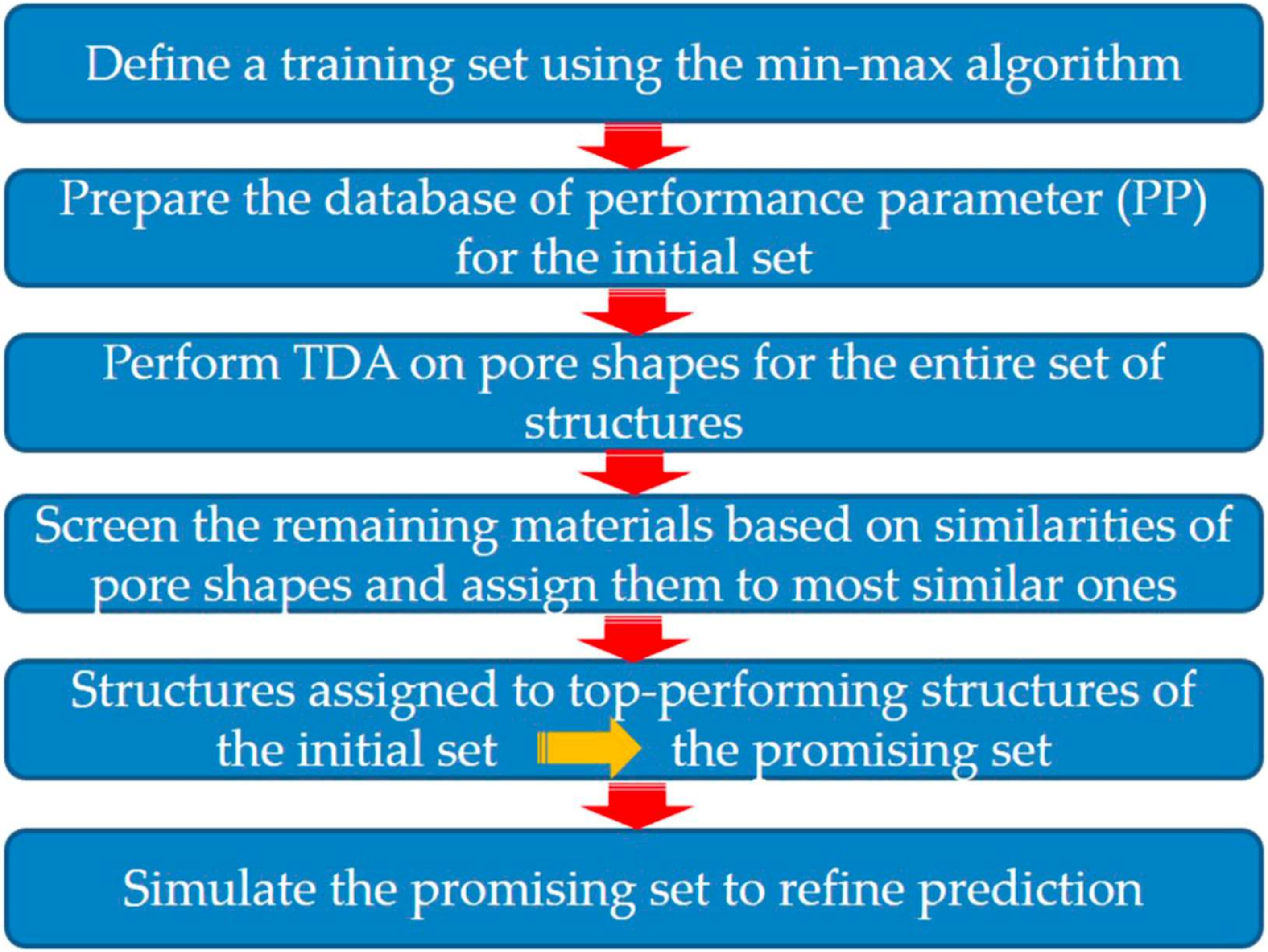


| | Examples | | | Features |
|---|---|---|---|---|
| Group A | **h8195264** | **h8196363** | **h8326835** | ● 1-dimensional channels<br>● Large equilateral polygonal cross-section |
| Group B | **h8053175** | **h8053791** | **h8101338** | ● 1-dimensional channels<br>● Small cross-section<br>● Multi channels |
| Group C | **h8196357** | **h8130054** | **h8193523** | ● 1-dimensional channels<br>● Un-equilateral polygonal cross-section |
| Group D | **h8123217** | **h8106667** | **h8202177** | ● Flatten channels<br>● Small void fraction, but average surface area |
| Group E | **h8193734** | **EMT** | **FAU** | ● Connected small polygonal cross-section<br>● Other shapes |
| Group F | **h8078190** | **h8059318** | **h8235214** | ● Small polygonal cross-section without connection |

# High-throughput screening


Nano-Porous Materials Genome
High-Throughput Screening
0
50
100
150
Pore Shape based Descriptor
Promising Structures

## Procedure of High-Throughput Screening



Define a training set using the min-max algorithm

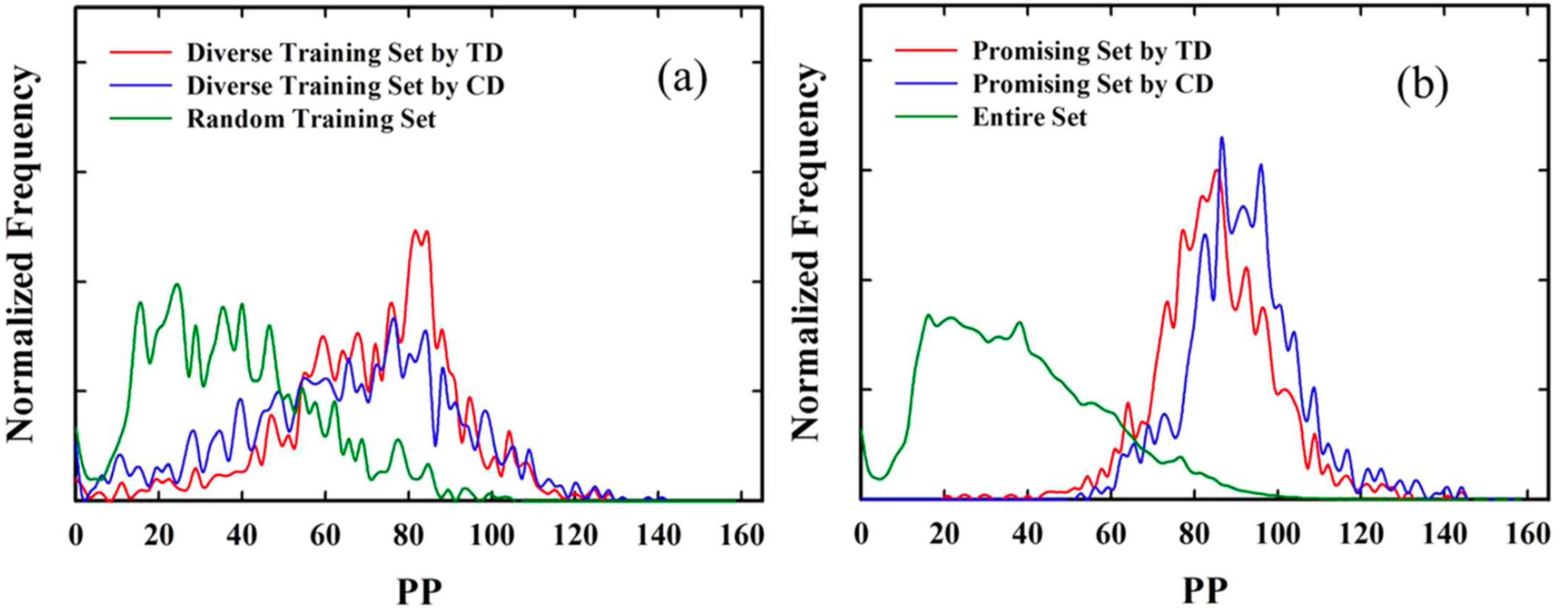
Prepare the database of performance parameter (PP) for the initial set

Perform TDA on pore shapes for the entire set of structures

Screen the remaining materials based on similarities of pore shapes and assign them to most similar ones

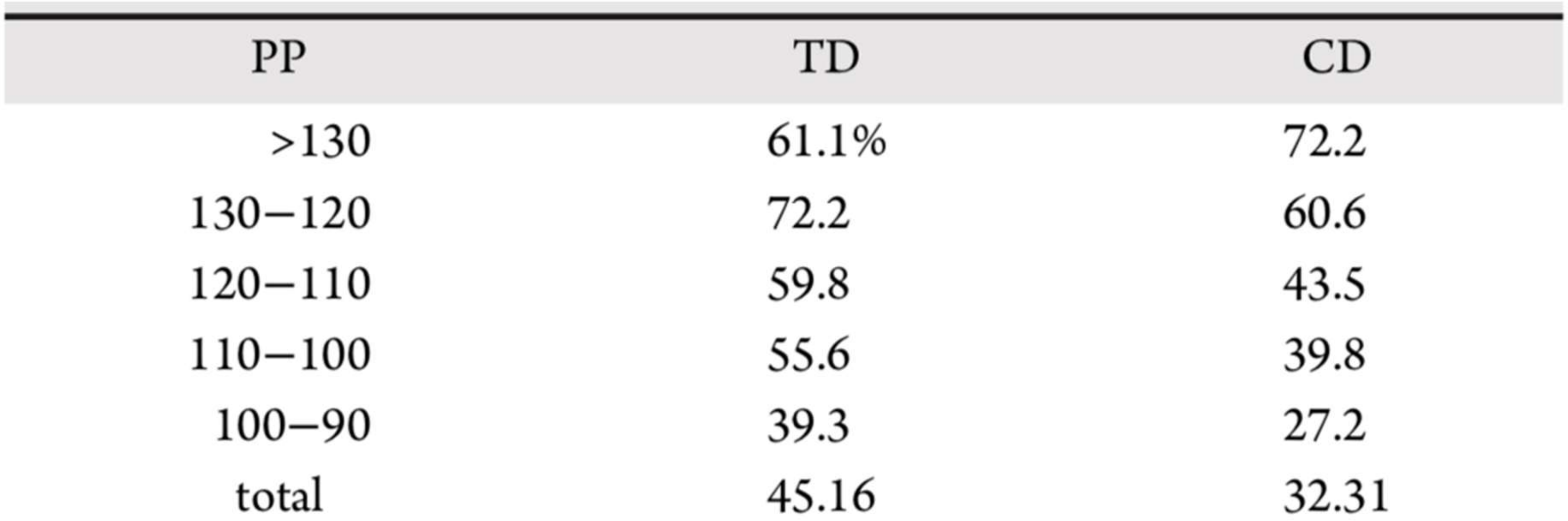
Structures assigned to top-performing structures of the initial set → the promising set

Simulate the promising set to refine prediction



(Normalized with respect to number of structures in each set.)

# Percentage of top 1% materials in promising sets

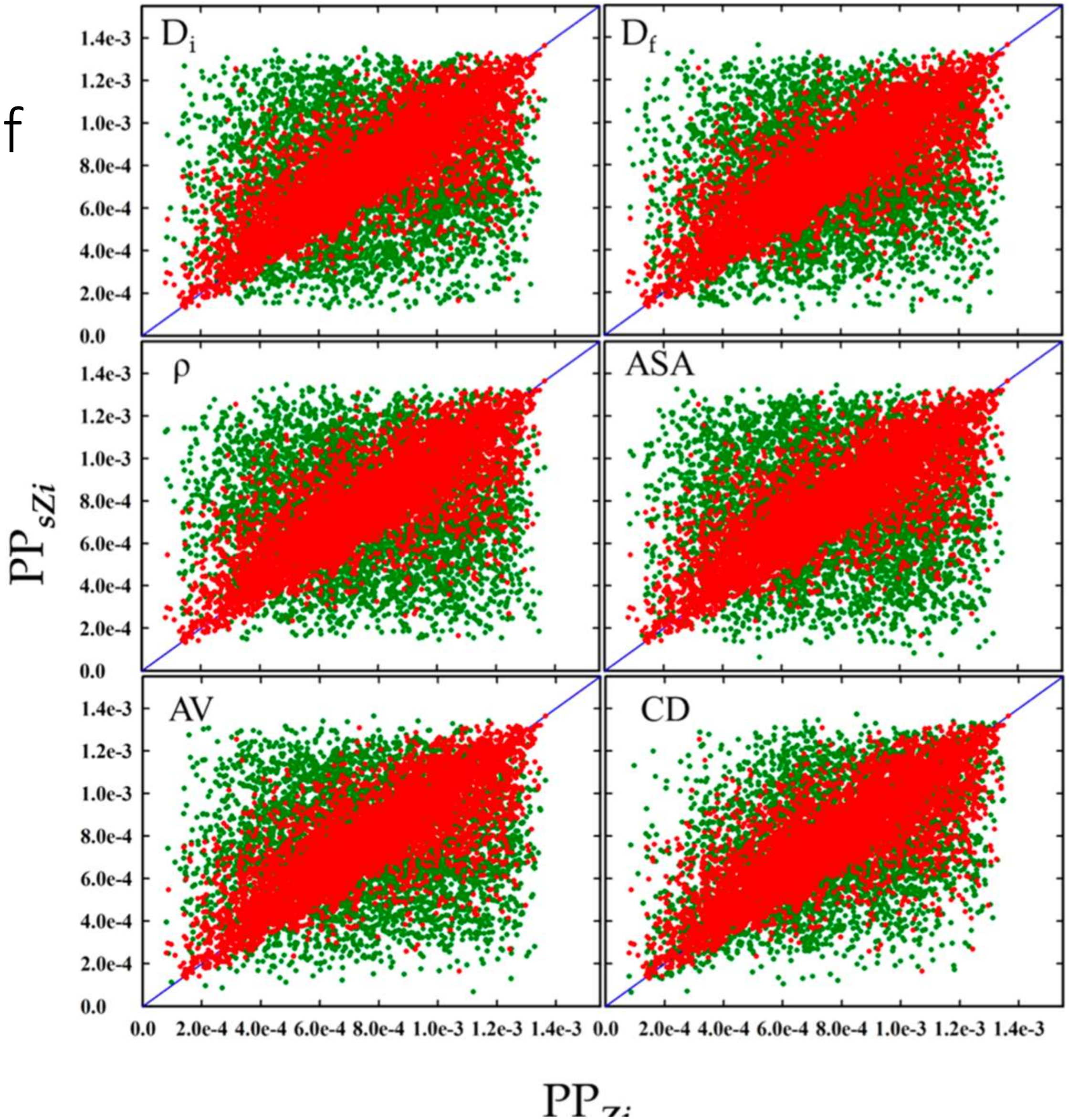


| PP | TD | CD |
|---|---|---|
| >130 | 61.1% | 72.2 |
| 130−120 | 72.2 | 60.6 |
| 120−110 | 59.8 | 43.5 |
| 110−100 | 55.6 | 39.8 |
| 100−90 | 39.3 | 27.2 |
| total | 45.16 | 32.31 |

# Another application: carbon capture

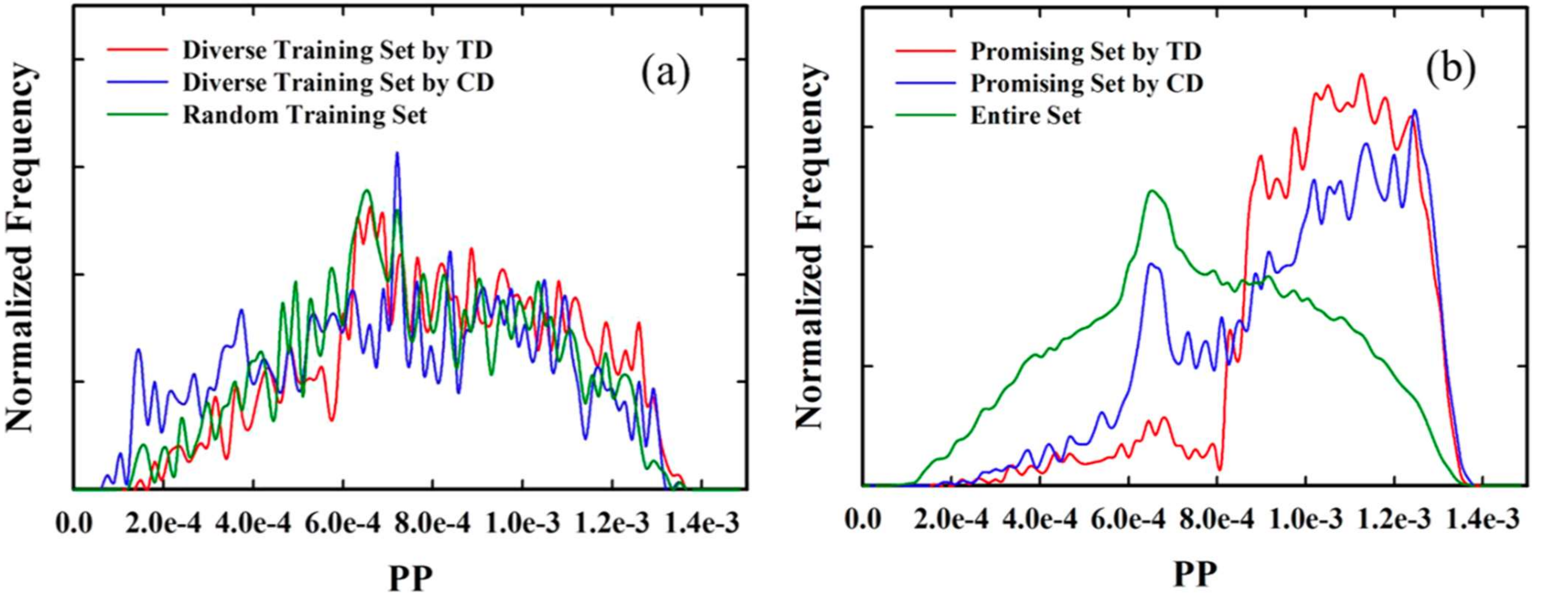
- **Relevant performance property:** *parasitic energy,* i.e., the loss of electricity production if a carbon capture-and-sequestration process is added to a coal-fired power plant.

# Parasitic energy of each zeolite vs its closest match



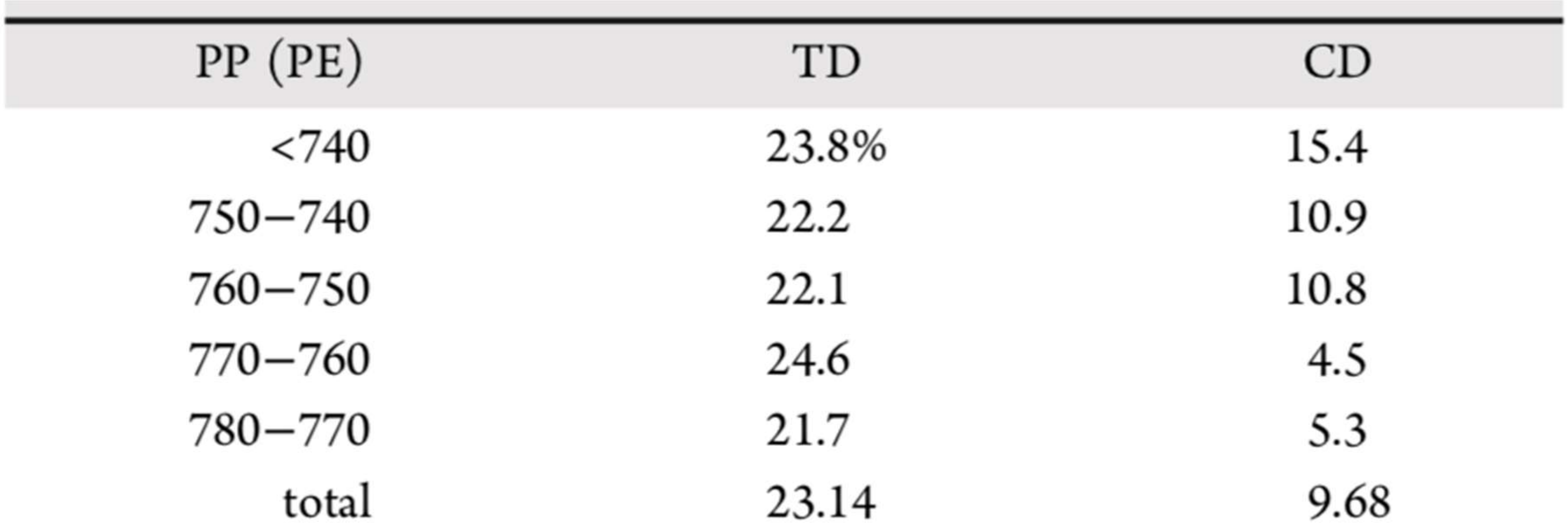
Red = TS

Green = classical



(Normalized with respect to the number of structures in each set.)

# Percentage of top 1% materials in promising sets



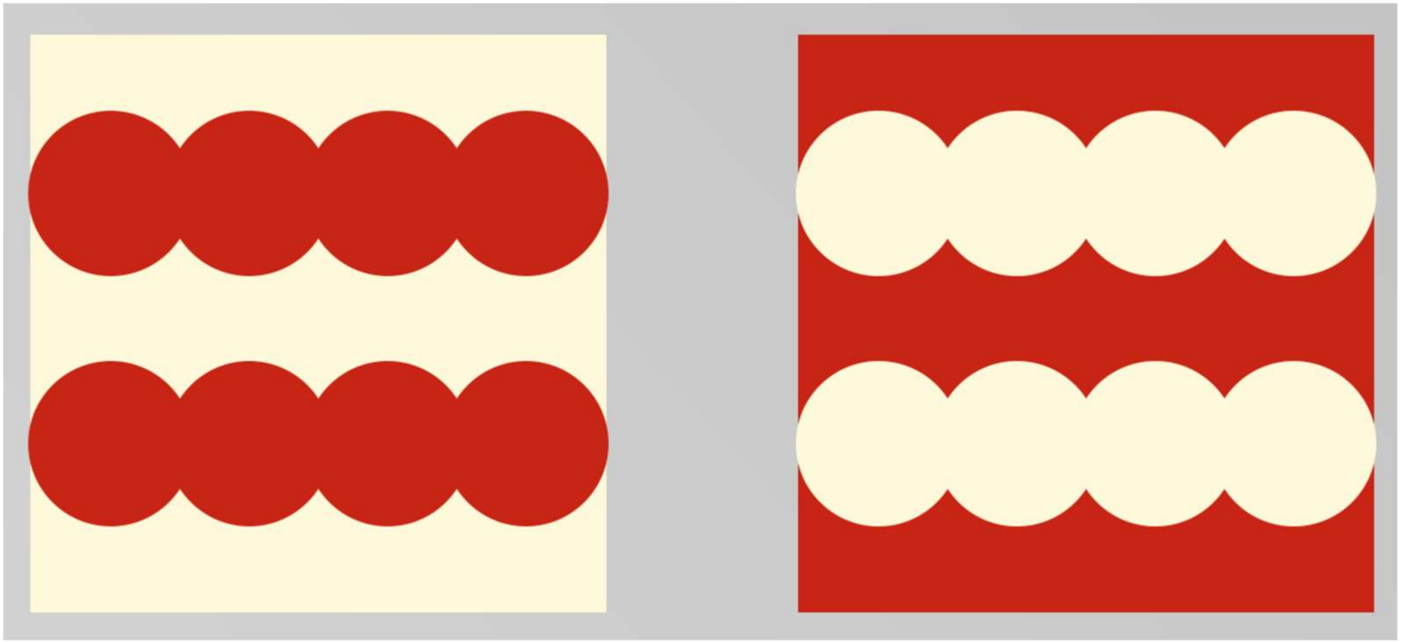
| PP (PE) | TD | CD |
|---|---|---|
| <740 | 23.8% | 15.4 |
| 750−740 | 22.2 | 10.9 |
| 760−750 | 22.1 | 10.8 |
| 770−760 | 24.6 | 4.5 |
| 780−770 | 21.7 | 5.3 |
| total | 23.14 | 9.68 |

# Conclusions

- Quantifying similarity of pore structures allows us not only to find structures geometrically similar to top-performing ones, but also to organize the set of materials with respect to the similarity of their pore shapes.
- For methane storage, we find several distinct classes of pore shapes and conclude that each class actually requires a different optimization strategy.
- In global searches for the most similar structures to a selected subset, the overall performance of TD is significantly better than that of the aggregate of CD.
- TD is highly capable of detecting good materials in the entire set, as long as some top materials are already known.
- The TD screening approach is highly efficient in detecting high-performing materials for both methane storage and carbon capture.

# Example of an open problem

Inside vs outside:



# Future work

- Use alpha complexes or cubical complexes, rather than Vietoris-Rips complexes.
- For applications in which the pores play a more active role, such as catalysis, extend the methodology to include chemical specificity and charge distribution.
- Find a way to take symmetries and periodic boundary conditions into account.
- Solve the “inside vs outside” problem.
- Find a way to deal with the fact that $D_{TS}$ currently too expensive compute for all possible pairs of predicted materials.

# Thank you!