HELMHOLTZ
MUNICH

AIH Institute of AI for Health

Zoom and Enhance
Towards Multi-Scale Representations in the Life Sciences
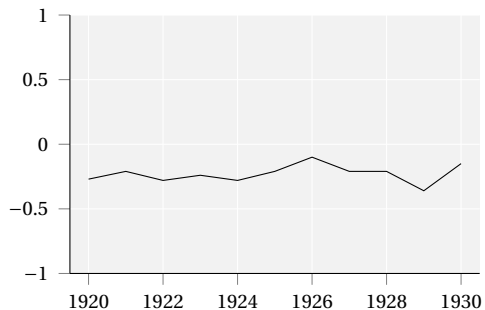Bastian Rieck (@Pseudomanifold)
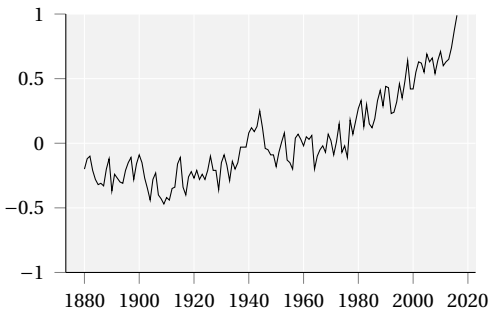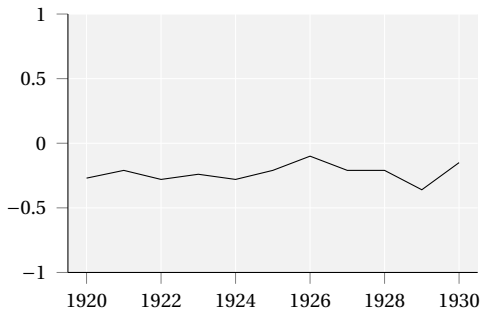
Motivation

Warm-up

It's all about finding the right perspective!

HELMHOLTZ MUNICH   AIH Institute of AI for Health

Warm-up

It's all about finding the right perspective!

Part I: Antimicrobial Resistance Prediction

Antimicrobial resistance

Relevance

☆ *It's time to confront the pandemic of antibiotic resistance*[1]

☆ *The World's Next Big Health Emergency Is Already Here*[2]

☆ *Millions are dying from drug-resistant infections, global report says*[3]

[1] https://www.ft.com/content/da746047-ecbc-4c4f-b95d-401421ce13c1
[2] https://www.bloomberg.com/opinion/articles/2022-01-27/after-covid-antimicrobial-resistance-is-the-world-s-biggest-health-emergency
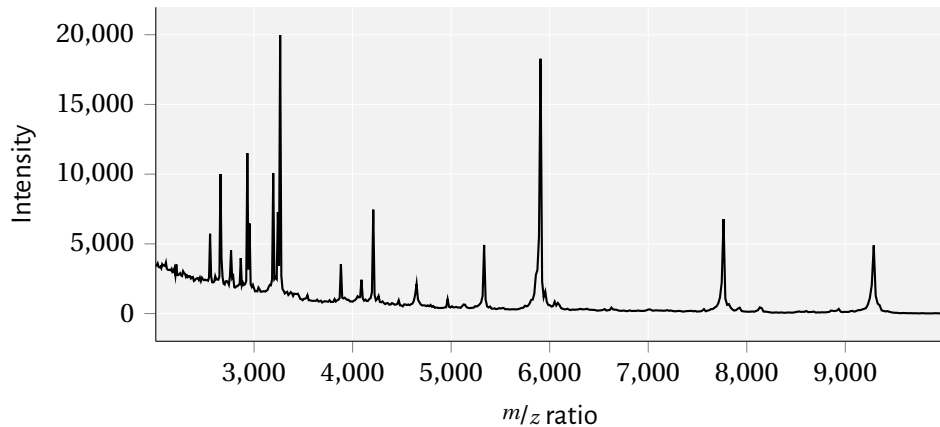[3] https://www.bbc.com/news/health-60058120

MALDI-TOF mass spectrometry

Matrix-**a**ssisted **l**aser **d**esorption **i**onisation **t**ime-**o**f-**f**light mass spectrometry
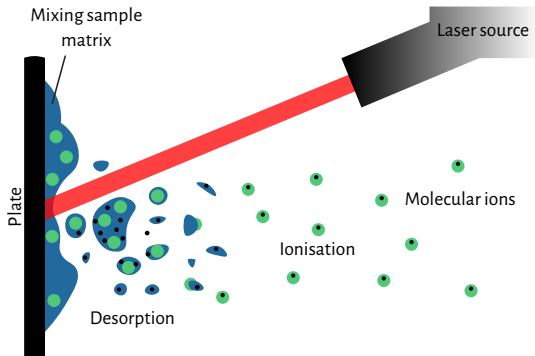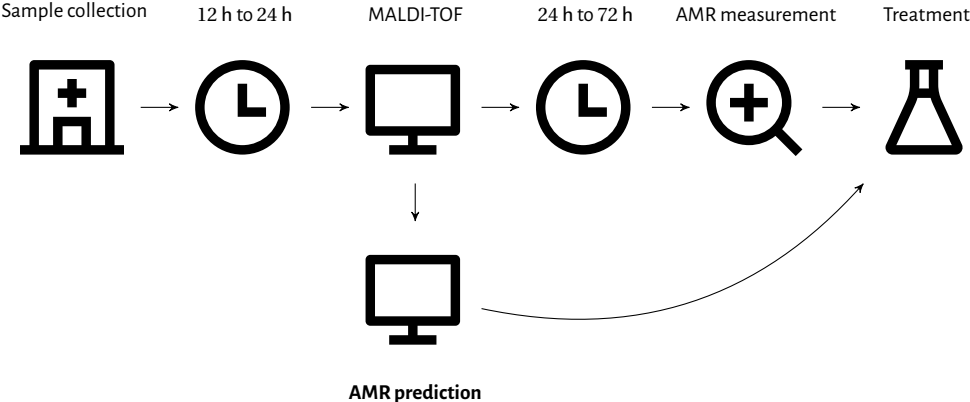

☆ Obtain a quick overview of sample microbial composition.

☆ Spectra are known to be *highly characteristic* of a microbial species.

☆ Standard tool for *species identification* in clinical practice.

Example

Antimicrobial treatment workflow

A tale of two pre-processing pipelines

1. Variance stabilisation
2. Smoothing
3. Baseline removal
4. Intensity calibration
5. Intensity trimming

State-of-the-art: S. Gibb and K. Strimmer, 'MALDIquant: a versatile R package for the analysis of mass spectrometry data', *Bioinformatics* 28.17, 2012, pp. 2270–2271

Treat spectrum as function and use the *prominence* of critical points as a proxy for the heights of a peak.

Digression

Critical points of a function

Prominence is also known as *topological persistence*, a concept from topological data analysis. See F. Hensel, M. Moor and **B. Rieck**, 'A Survey of Topological Machine Learning Methods', *Frontiers in Artificial Intelligence* 4, 2021 for a recent survey.
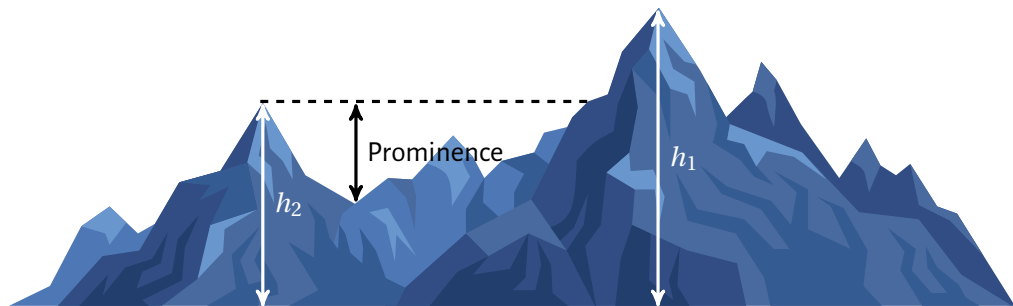
Digression

Critical points of a function

Prominence is also known as *topological persistence*, a concept from topological data analysis. See F. Hensel, M. Moor and **B. Rieck**, 'A Survey of Topological Machine Learning Methods', *Frontiers in Artificial Intelligence* 4, 2021 for a recent survey.
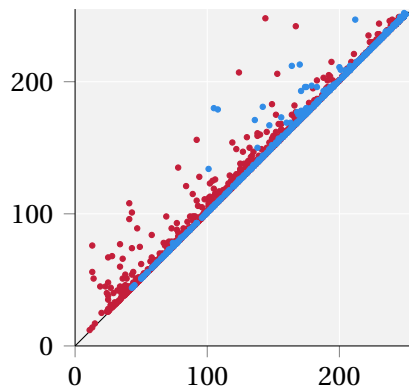
Digression

Critical points of a function

Prominence is also known as *topological persistence*, a concept from topological data analysis. See F. Hensel, M. Moor and **B. Rieck**, 'A Survey of Topological Machine Learning Methods', *Frontiers in Artificial Intelligence* 4, 2021 for a recent survey.
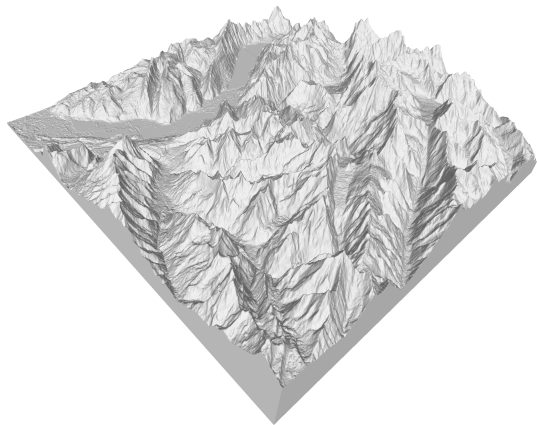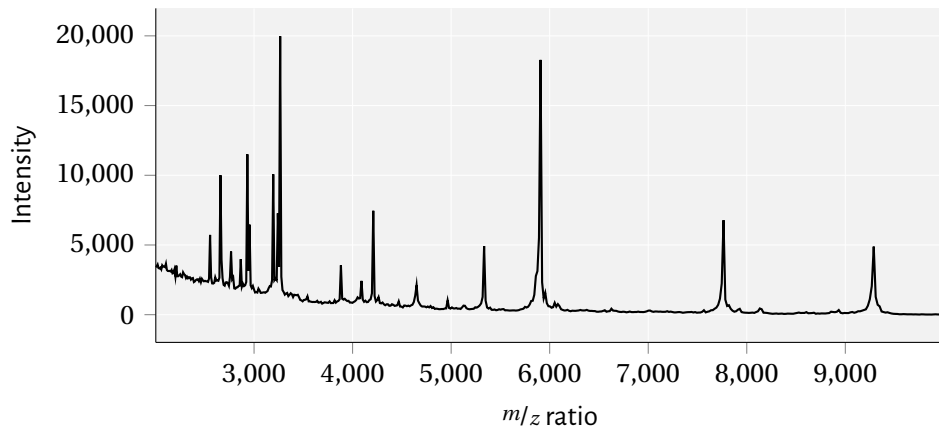
Digression

Higher dimensions

The calculation of topological features is also known as *persistent homology*. It results in a set of topological descriptors, the *persistence diagrams*.
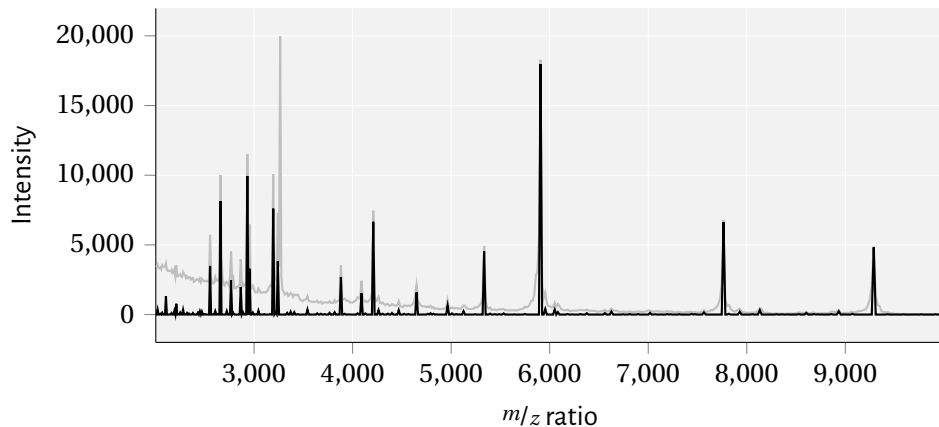
Example

Topological pre-processing

Example

Topological pre-processing

Using spectra for machine learning tasks

☆ Choose bin size to obtain fixed-sized feature vectors
☆ Use 'sparse' representation based on tuples

The latter part is reminiscent of M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov and A. J. Smola, 'Deep Sets', *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 3391–3401.

Towards antimicrobial resistance prediction

PIKE: Peak Information Kernel

For two spectra S and S', with m/z values x_i and x'_i and intensities λ_i and λ'_i, respectively, we calculate the following expression:

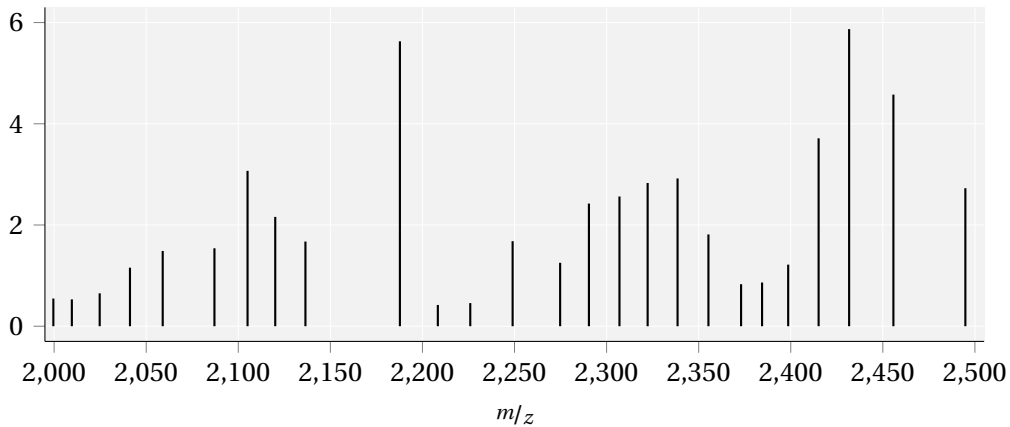$$k_t(S, S') = \frac{1}{2\sqrt{2\pi t}} \sum_{i,j} \lambda_i \lambda'_j \exp\left(-\frac{\left(x_i - x'_j\right)^2}{8t}\right)$$

Properties

- ☆ C^∞ function with closed-form gradient expression
- ☆ Calculate similarity based on peak 'distances'
- ☆ Interactions between peaks are captured
- ☆ Single parameter $t \in \mathbb{R}$ controls smoothing
- ☆ Can be easily integrated into *any* kernel-based model: SVM, Gaussian Processes, …

Example

Varying t
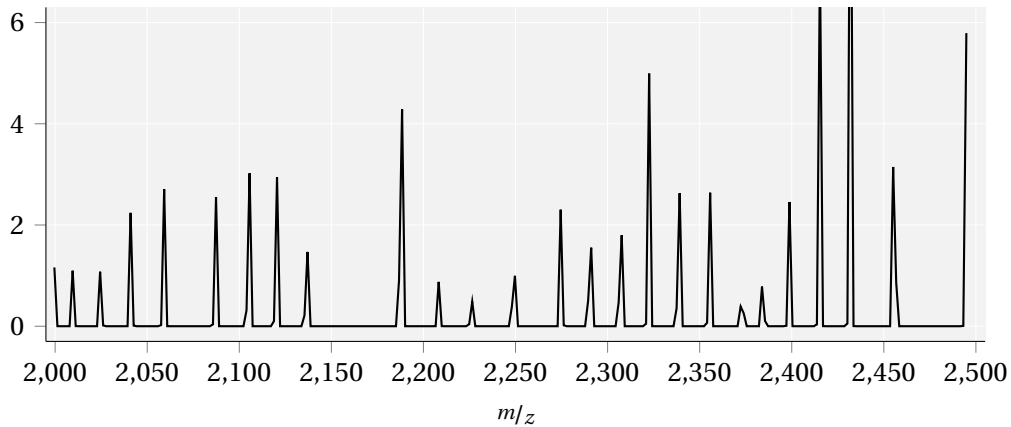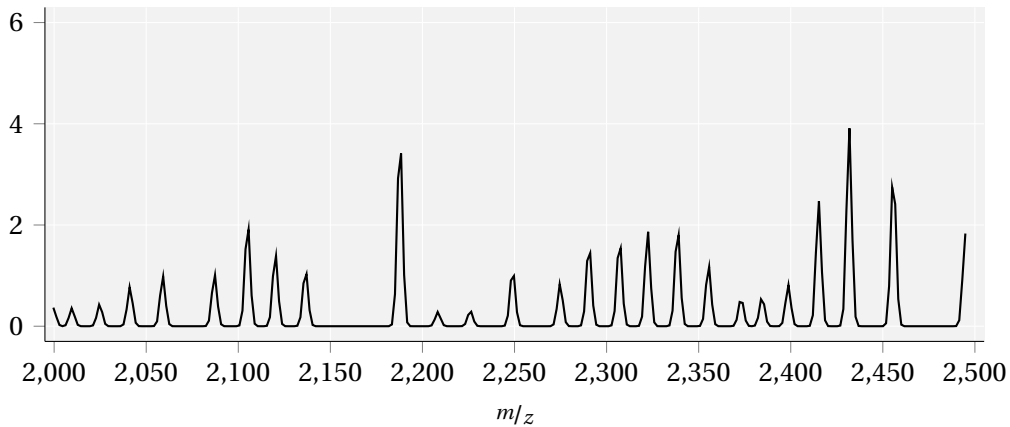

$t \approx 0$

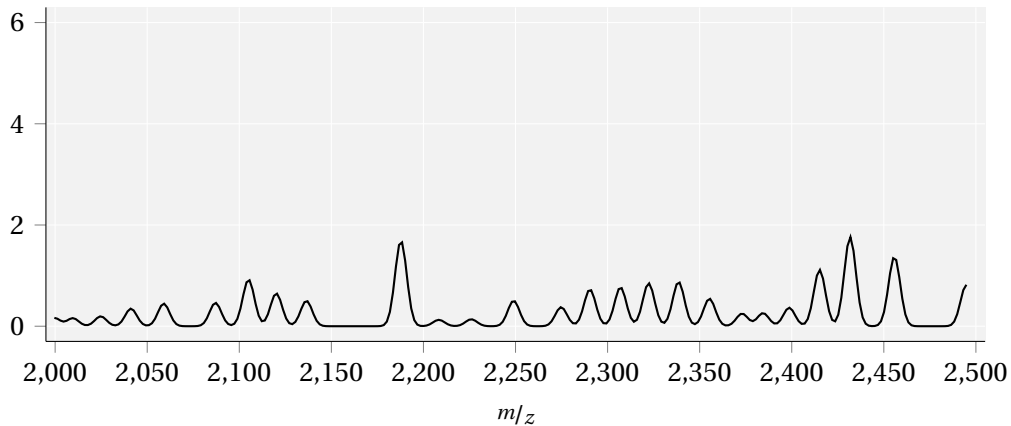# Example

Varying t

$t = 0.10$

Example

Varying t

$t = 1$

Example

Varying t

$t = 5$

HELMHOLTZ MUNICH   AIH Institute of AI for Health

Example

Varying t

$t = 10$
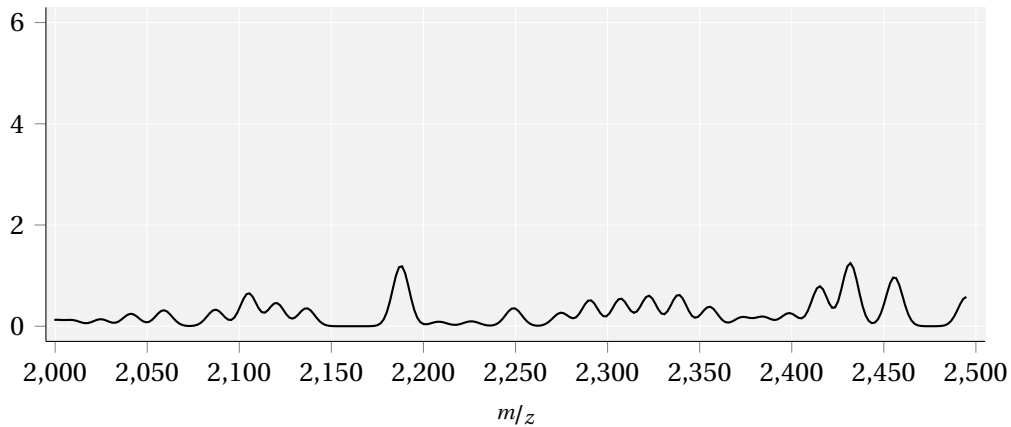
Example

Varying t

$t = 50$

Example

Varying t

$t = 100$

HELMHOLTZ MUNICH AIH Institute of AI for Health
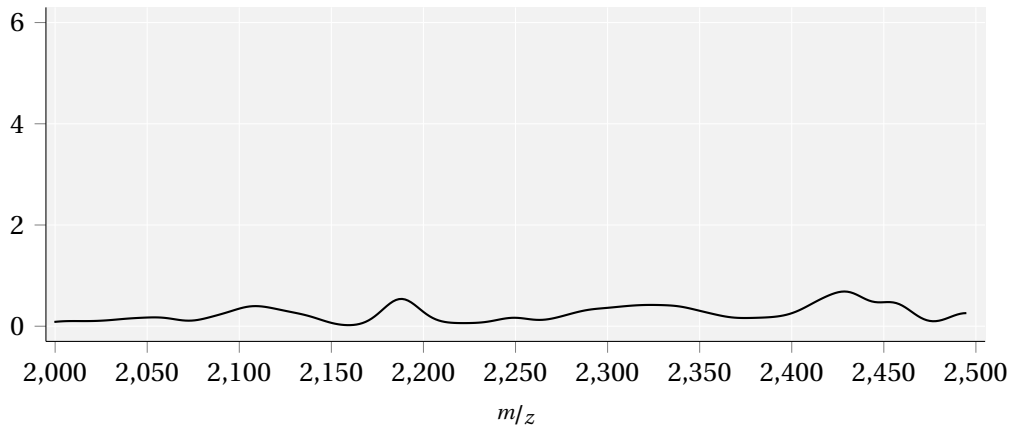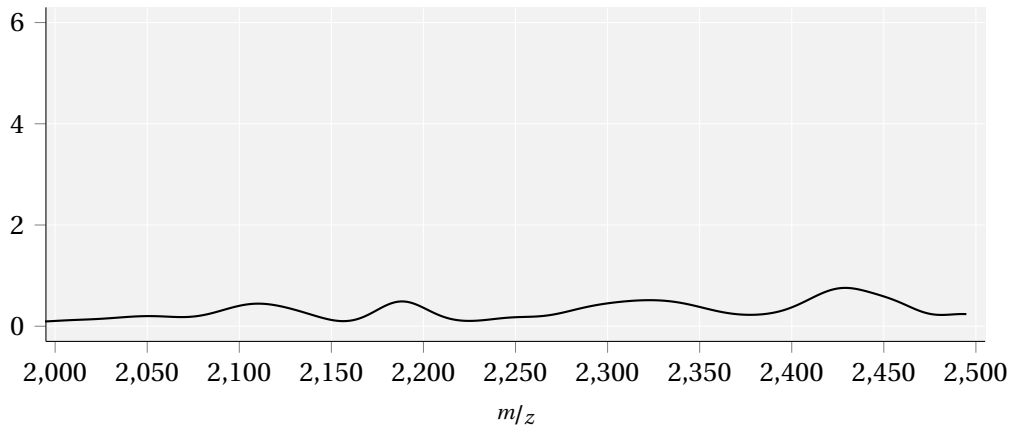
Data set

Species	Antibiotic	Samples	% resistant
E. coli	amoxicillin / clavulanic acid	1043	28.9
	ceftriaxone	1060	20.4
	ciprofloxacin	1051	29.7
K. pneumoniae	ceftriaxone	597	15.1
	ciprofloxacin	596	16.8
	piperacillin / tazobactam	576	13.9
S. aureus	amoxicillin / clavulanic acid	973	13.7
	ciprofloxacin	987	14.7
	penicillin	941	71.4

Publication

C. Weis*, M. Horn*, **B. Rieck**, A. Cuénod, A. Egli and K. Borgwardt, 'Topological and kernel-based microbial phenotype prediction from MALDI-TOF mass spectra', *Bioinformatics* 36.Supplement_1, 2020, pp. i30–i38

HELMHOLTZ MUNICH   AIH Institute of AI for Health

GP-PIKE: superior performance

AUPRC

Species	Antibiotic	LR	GP–RBF	GP–PIKE
E. coli	amox/clav[1]	41.0 ± 7.4	32.5 ± 8.5	$\mathbf{47.1 \pm 3.9}$
	ceftriaxone	63.2 ± 6.1	46.3 ± 24.0	$\mathbf{70.6 \pm 3.2}$
	ciprofloxacin	61.4 ± 8.5	34.7 ± 10.7	$\mathbf{68.0 \pm 3.0}$
K. pneumoniae	ceftriaxone	58.2 ± 9.8	58.7 ± 25.3	$\mathbf{77.0 \pm 6.8}$
	ciprofloxacin	41.7 ± 9.8	30.9 ± 13.5	$\mathbf{54.6 \pm 10.1}$
	pip/tazo[2]	31.6 ± 6.8	13.8 ± 0.0	$\mathbf{56.5 \pm 9.7}$
S. aureus	amox/clav[1]	52.9 ± 3.9	13.9 ± 0.0	$\mathbf{69.2 \pm 9.2}$
	ciprofloxacin	34.1 ± 3.3	23.3 ± 11.9	$\mathbf{39.4 \pm 6.6}$
	penicillin	79.7 ± 3.3	74.2 ± 3.2	$\mathbf{83.2 \pm 3.5}$

Advantages & disadvantages

Sparse processing and 'built-in' confidence analysis, but insufficient scalability to larger data sets. How does a model fare on larger data sets?

Bigger and better?

- ☆ 303,195 mass spectra
- ☆ 768,300 antimicrobial resistance labels
- ☆ 803 different species of bacterial and fungal pathogens
- ☆ 4 different diagnostic laboratories

Focus on *scalability* first: only using 'standard' classifiers based on 6000-dimensional feature vectors, obtained from binning the spectra.

Publication

C. Weis, A. Cuénod, **B. Rieck**, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff, K. Borgwardt[†] and A. Egli[†], 'Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning', *Nature Medicine* 28, 2022, pp. 164–174

Data

`https://doi.org/10.5061/dryad.bzkh1899q`
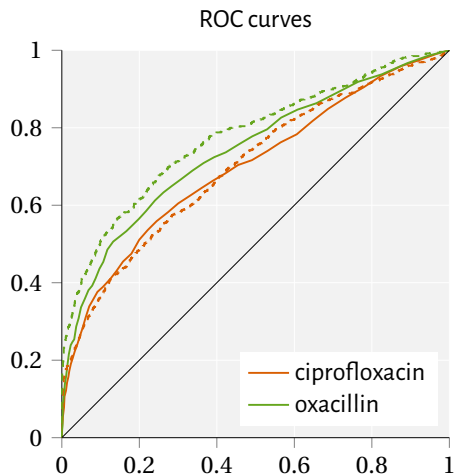
Results

E. coli (excerpt)

ROC curves

Precision–recall curves

ceftriaxone
ciprofloxacin
pip/tazo
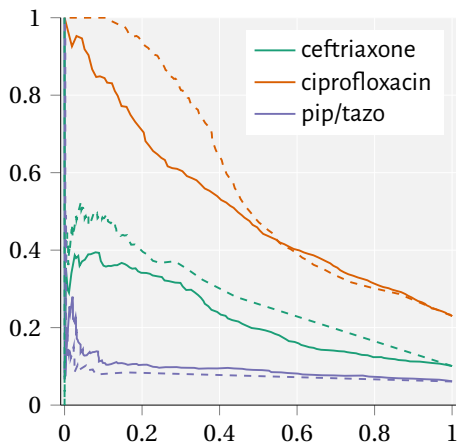
(solid: LR, dashed: LightGBM)

Results

S. aureus (excerpt)

ROC curves

Precision–recall curves

(solid: LR, dashed: LightGBM)
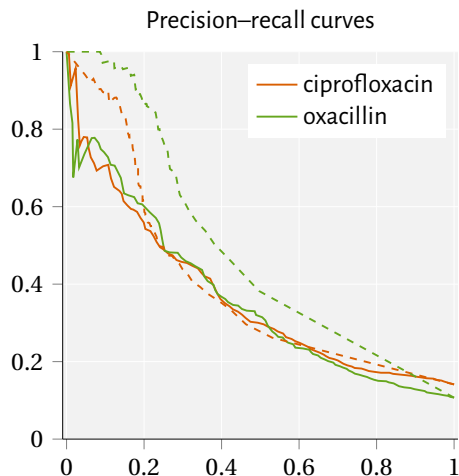
Going to other sites

E. coli, AUROC

	DRIAMS-A	DRIAMS-B	DRIAMS-C	DRIAMS-D
DRIAMS-A	0.74	0.81	0.64	0.68
DRIAMS-B	0.59	0.71	0.59	0.59
DRIAMS-C	0.58	0.62	0.66	0.59
DRIAMS-D	0.63	0.76	0.66	0.75

K. pneumoniae, AUROC

	DRIAMS-A	DRIAMS-B	DRIAMS-C	DRIAMS-D
DRIAMS-A	0.74	0.44	0.67	0.61
DRIAMS-B	0.53	0.57	0.51	0.61
DRIAMS-C	0.53	0.53	0.74	0.63
DRIAMS-D	0.61	0.39	0.54	0.71

What is the structure of spectra from different sites?

Preliminary work
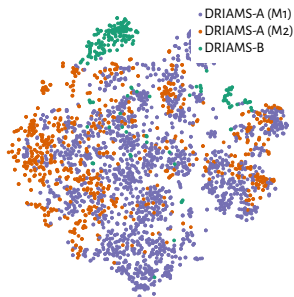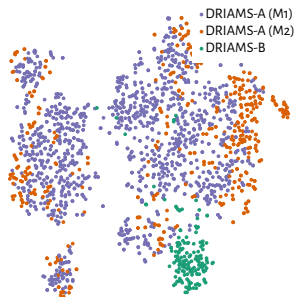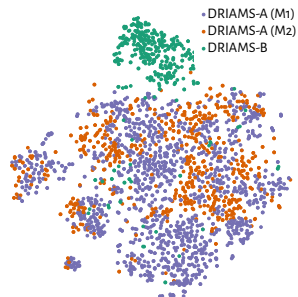
Different *confounding effects* at work here!

Lessons learned

   ✩ 'Relatively simple' models (LR, LightGBM, MLP) already exhibit useful performance

   ✩ Calibration of classifiers is necessary to support rejecting samples

   ✩ Feature importance highlights interesting peaks in a spectrum

Challenges

   ✩ Domain adaptation

   ✩ Extracting features that *generalise* over time

Part II: Topology-Driven fMRI Data Analysis

fMRI data

Our approach

- ✩ Consider the BOLD activation function f to be a time-varying function on a manifold $\mathcal{M}$
- ✩ Calculate topological features of $\mathcal{M}$ 'measured' via f
- ✩ Obtain stable topological summaries at different resolutions

Main advantage of this approach

Working on the 'raw' data; no auxiliary representations necessary! In particular, no *atlas* required (fewer modelling choices in total).

Publication

B. Rieck* et al., 'Uncovering the Topology of Time-Varying fMRI Data using Cubical Persistence', *Advances in Neural Information Processing Systems*, vol. 33, Accepted as a *spotlight* presentation at NeurIPS (**top 3%** of all submissions), 2020, pp. 6900–6912, arXiv: `2006.07882 [q-bio.NC]`

Workflow
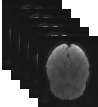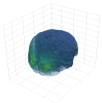
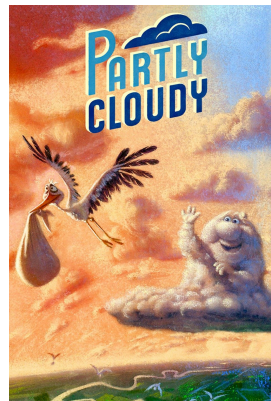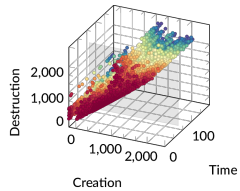| fMRI stack | fMRI volume | Cubical complex | Persistence diagram | Persistence images |
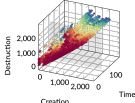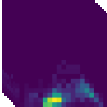
Our data set

- 155 (122 children, 33 adults) participants are being shown the film 'Partly Cloudy'
- *Continuous* stimulation of participants
- 168 time steps
- *No additional information about participants has been provided on purpose*

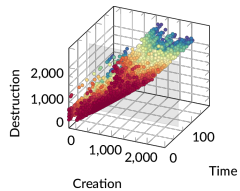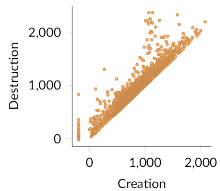# Topological summaries

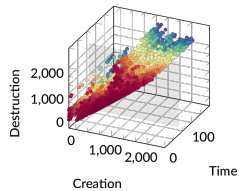Persistence diagrams
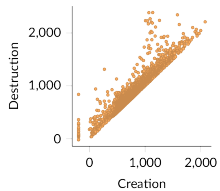
Topological summaries

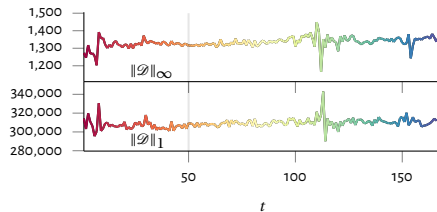Persistence diagrams

Persistence diagram

Topological summaries

Persistence diagrams

Persistence diagram

Summary statistics

Age prediction based on summary statistics

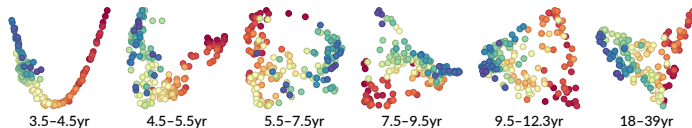| Method	BM	OM	XM
baseline-tt	0.09	0.02	0.24
baseline-pp	0.41	0.40	0.40
tt-corr-tda	0.17	0.11	0.23
pp-corr-tda	0.25	0.27	0.23
srm	0.44		
$\|\mathscr{D}\|_1$	0.46	0.67	0.48
$\|\mathscr{D}\|_\infty$	**0.61**	**0.77**	**0.73**

Brain state trajectories
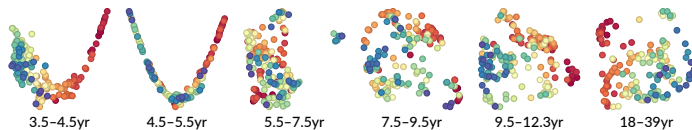
Whole-brain mask
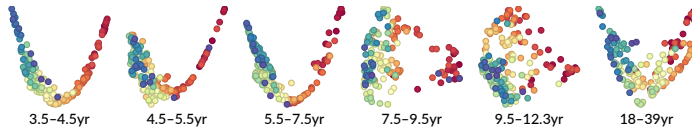
Occipital-temporal mask

XOR mask

Summary

- ☆ Topological features capture salient information about participants
- ☆ Want to extend this to larger and more complex data sets
- ☆ Interesting challenge: how to obtain noise-impervious representations?

Topology can provide a useful set of inductive biases for uncovering salient features at different resolutions, *without* imposing strong restrictions on the representation of the data.

Acknowledgements