ETH zürich

Machine learning for exploring biological systems

Keynote

Karsten Borgwardt

ETH Zürich, D-BSSE

Turkish Science Academy, June 23, 2021

Machine learning and systems biology

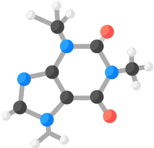
Goals

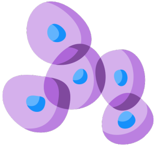
- Machine learning tries to detect statistical dependencies in large datasets.

Machine learning and systems biology

Goals

- Machine learning tries to detect statistical dependencies in large datasets.





- Systems biology studies the interplay of components of a biological system and the functions/properties it gives rise to.

Machine learning and systems biology

Motivation

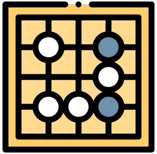
- Enormous success of machine learning in tasks such as classifying images, recognizing speech, translating text, and playing games

Machine learning and systems biology

Motivation

- Enormous success of machine learning in tasks such as classifying images, recognizing speech, translating text, and playing games



- Can this success be translated to systems biology, and the life sciences in general?

Machine learning and systems biology

Holy grails of computational biology

- Structural biology: predicting protein structure from protein sequence
- Genetics: predicting complex traits of individuals based on their genotypes

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

Machine learning and systems biology

Further central topics

- Chemoinformatics: predicting function based on molecular structure

Machine learning and systems biology

Further central topics

- Chemoinformatics: predicting function based on molecular structure
- Medicine: predicting disease diagnosis, progression, therapy outcome

Machine learning and systems biology

Further central topics

- Chemoinformatics: predicting function based on molecular structure
- Medicine: predicting disease diagnosis, progression, therapy outcome
- Genomics: predicting e.g. the exact position of a gene within the genome

Machine learning and systems biology

Further central topics

- Chemoinformatics: predicting function based on molecular structure
- Medicine: predicting disease diagnosis, progression, therapy outcome
- Genomics: predicting e.g. the exact position of a gene within the genome

Common problem: insufficient prediction accuracy

Machine learning and systems biology

Obstacles for machine learning in the life sciences

1. Not enough observations

Machine learning and systems biology

Obstacles for machine learning in the life sciences

1. Not enough observations
2. Uncertainty and difficulty in phenotyping

Machine learning and systems biology

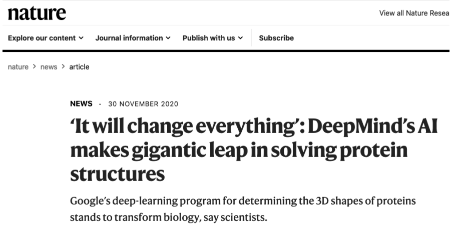
Obstacles for machine learning in the life sciences

1. Not enough observations
2. Uncertainty and difficulty in phenotyping
3. Unclear which complexity of machine learning models is required

Machine learning and systems biology

Recently big progress

- Protein structure prediction



nature

View all Nature Resea

Explore our content · Journal information · Publish with us · Subscribe

nature > news > article

NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

- Molecular function prediction

Science that inspires

Cell

ARTICLE | VOLUME 180, ISSUE 4, P688-702.E13, FEBRUARY 20, 2020

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes • Kevin Yang [10] • Kyle Swanson [10] • ... Tommi S. Jaakkola • Regina Barzilay • James J. Collins [11] • Show all authors • Show footnotes

DOI: https://doi.org/10.1016/j.cell.2020.01.021 • Check for updates

Machine learning and systems biology

Recently big progress

- Protein structure prediction

nature

View all Nature Resea

Explore our content ⌄ Journal information ⌄ Publish with us ⌄ Subscribe

nature › news › article

NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

- Molecular function prediction

Science that inspires

Cell

ARTICLE | VOLUME 180, ISSUE 4, P688-702.E13, FEBRUARY 20, 2020

A Deep Learning Approach to Antibiotic Discovery

Jonathan M. Stokes • Kevin Yang [10] • Kyle Swanson [10] • ... Tommi S. Jaakkola • Regina Barzilay • James J. Collins [11] • Show all authors • Show footnotes

DOI: https://doi.org/10.1016/j.cell.2020.01.021 • Check for updates

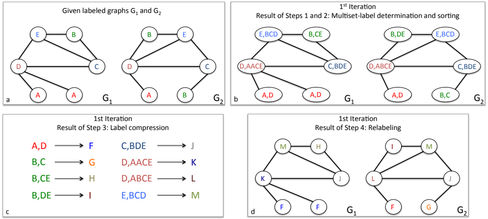
Both use machine learning on graphs

Machine learning on graphs

Machine learning and systems biology

Machine learning on graphs

- Graphs are the data structure to represent systems, networks and structures.

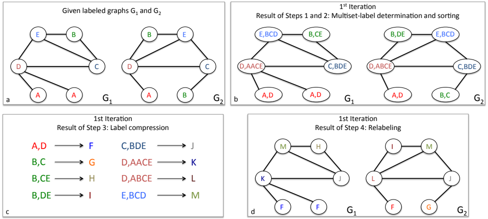


Shervashidze et al., 2011

Machine learning and systems biology

Machine learning on graphs

- Graphs are the data structure to represent systems, networks and structures.
- Graph comparison in practice computationally expensive (Borgwardt et al., 2005)

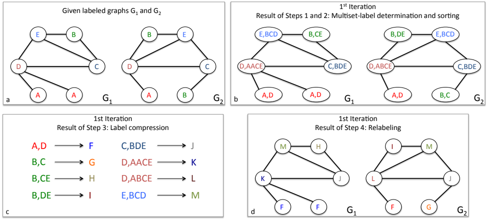


Shervashidze et al., 2011

Machine learning and systems biology

Machine learning on graphs

- Graphs are the data structure to represent systems, networks and structures.
- Graph comparison in practice computationally expensive (Borgwardt et al., 2005)
- Fast *graph kernels* based on the Weisfeiler-Lehman scheme (Shervashidze and Borgwardt, 2009; Shervashidze et al., 2011)

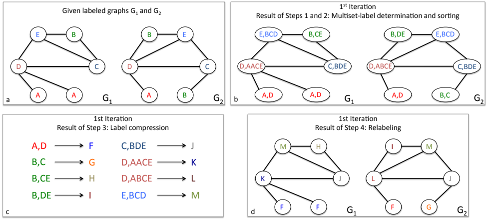


Shervashidze et al., 2011

Machine learning and systems biology

Machine learning on graphs

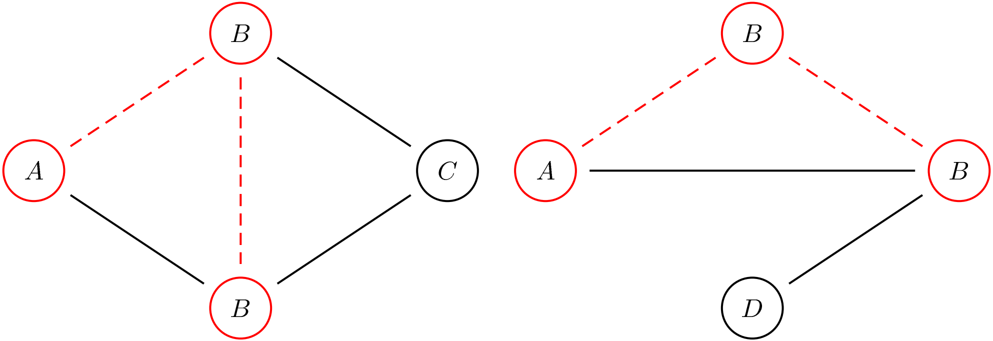
- Graphs are the data structure to represent systems, networks and structures.
- Graph comparison in practice computationally expensive (Borgwardt et al., 2005)
- Fast *graph kernels* based on the Weisfeiler-Lehman scheme (Shervashidze and Borgwardt, 2009; Shervashidze et al., 2011)
- Fundamental concept in *graph kernels* and *graph convolutional networks* (Borgwardt et al., Foundations and Trends in Machine Learning 2020)



Shervashidze et al., 2011

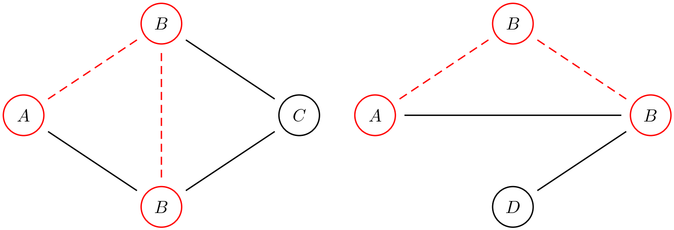
Machine learning on graphs

Fundamental question: How similar are two graphs?



Machine learning on graphs

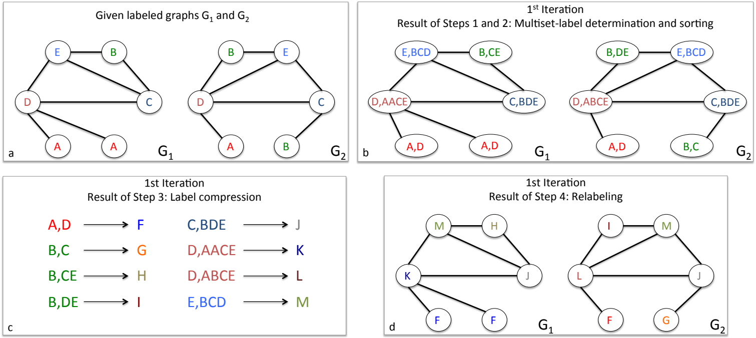
1. Similarity measures on graphs: Counting matching subgraphs



- Basis of many past and current graph representations, e.g.:
  - random walk kernels (Kashima et al., 2003 and Gärtner et al., 2003)
  - shortest paths kernels (Borgwardt and Kriegel, 2005)
  - graphlets (Przulj, 2007)

Machine learning on graphs

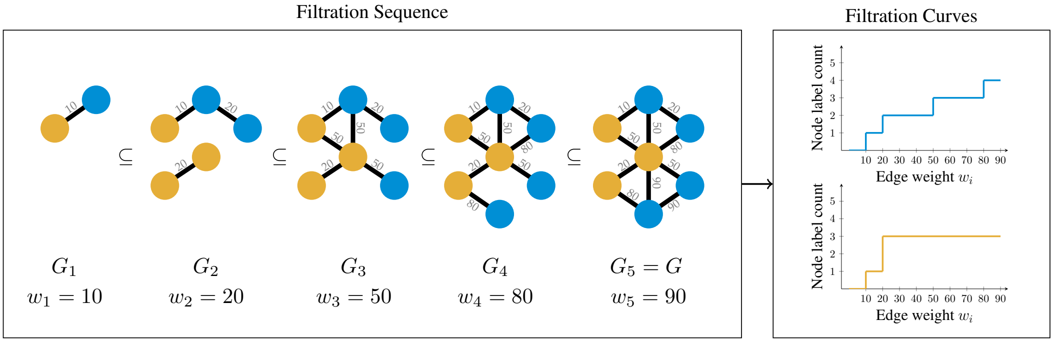
2. Similarity measures on graphs: Neighborhood aggregation



- Basis of Weisfeiler-Lehman graph kernels and (Spatial) Graph Convolutional Networks (e.g., Shervashidze et al., 2009, 2011, Kipf et al., 2016)

Machine learning on graphs

New graph representation approach: Filtration curves (O'Bray*, Rieck*, B., KDD 2021)



Machine learning on graphs

Filtration curve representation

Two components:

1. A graph filtration $\mathcal{F}_G$

- (native) edge weight
- max-degree
- Ricci curvature
- Heat kernel signature

Machine learning on graphs

Filtration curve representation

Two components:

1. A graph filtration $\mathcal{F}_G$

- (native) edge weight
- max-degree
- Ricci curvature
- Heat kernel signature

2. A graph descriptor function f

- Node label histogram
- Count of connected components

Machine learning on graphs

Filtration curve representation

Two components:

1. A graph filtration $\mathcal{F}_G$

- (native) edge weight
- max-degree
- Ricci curvature
- heat kernel signature

2. A graph descriptor function f

- node label histogram
- count of connected components

Machine learning on graphs

Filtration curve representation

Two components:

1. A graph filtration $\mathcal{F}_G$

- (native) edge weight
- max-degree
- Ricci curvature
- heat kernel signature

2. A graph descriptor function f

- node label histogram
- count of connected components

Runtime: $O(m \log m)$ for sorting all m edges

Machine learning on graphs

Filtration-based graph representation

- Given
  - a *graph filtration* $\mathcal{F}_G = (G_1, \ldots, G_m)$.
  - and a *graph descriptor function* $f : \mathcal{G} \to \mathbb{R}^d$

Then we can represent G as a high-dimensional *path* via

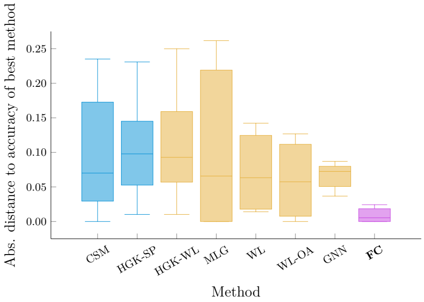
$$\mathcal{P}_G := \bigoplus_{i=1}^{m} f(G_i) \in \mathbb{R}^{m \times d}, \tag{1}$$

- where
  - m indexes the number of edge weight thresholds in $\mathcal{F}_G$, and
  - $\oplus$ refers to the concatenation operator.

Machine learning on graphs

Empirical comparison

- **Setup**: subgraph enumeration (blue) and neighborhood-aggregation (yellow) approaches versus Filtration Curves (pink) on graph classification benchmarks
- **Datasets**: collection of 8 labeled and 5 unlabeled datasets for graph classification



Machine learning on graphs

Filtration curves

- Efficient to compute and expressive graph representation
  - Code: `https://github.com/BorgwardtLab/filtration_curves`
  - General graph kernel code (Sugiyama et al., Bioinformatics 2018)

Machine learning on graphs

Filtration curves

- Efficient to compute and expressive graph representation
  - Code: `https://github.com/BorgwardtLab/filtration_curves`
  - General graph kernel code (Sugiyama et al., Bioinformatics 2018)

Impact of learning on graphs

- Growing number of successful applications in systems and network biology (Muzio*, O'Bray* et al., Briefings in Bioinformatics 2021)

Machine learning on graphs

Filtration curves

- Efficient to compute and expressive graph representation
  - Code: `https://github.com/BorgwardtLab/filtration_curves`
  - General graph kernel code (Sugiyama et al., Bioinformatics 2018)

Impact of learning on graphs

- Growing number of successful applications in systems and network biology (Muzio*, O'Bray* et al., Briefings in Bioinformatics 2021)
- Numerous further topics beyond graph comparison: e.g., graph generation and its evaluation (O'Bray et al., arXiv 2021 `https://arxiv.org/abs/2106.01098`)

Machine learning on graphs

Filtration curves

- Efficient to compute and expressive graph representation
  - Code: `https://github.com/BorgwardtLab/filtration_curves`
  - General graph kernel code (Sugiyama et al., Bioinformatics 2018)

Impact of learning on graphs

- Growing number of successful applications in systems and network biology (Muzio*, O'Bray* et al., Briefings in Bioinformatics 2021)
- Numerous further topics beyond graph comparison: e.g., graph generation and its evaluation (O'Bray et al., arXiv 2021 `https://arxiv.org/abs/2106.01098`)
- Inherently related to learning on sequences, time series and images - which also have manifold (potential) applications in the life sciences

Machine learning and systems biology

Example of success

- Synthetic biology: ribosome binding site (RBS) activity prediction

Machine learning and systems biology

Example of success

- Synthetic biology: ribosome binding site (RBS) activity prediction

Examples of ongoing work

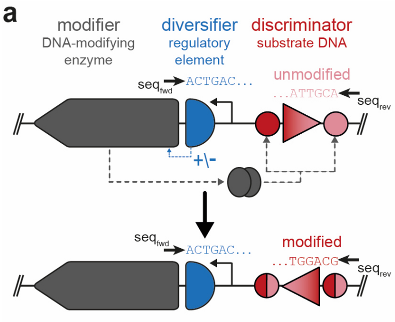
- Medicine: Sepsis prediction
- Plant breeding: Wheat yield prediction

Machine learning in synthetic biology

Ribosome binding site activity prediction

DNA-based phenotypic recording (Höllerer*, Papaxanthos*, et al., Nature Comm 2020)

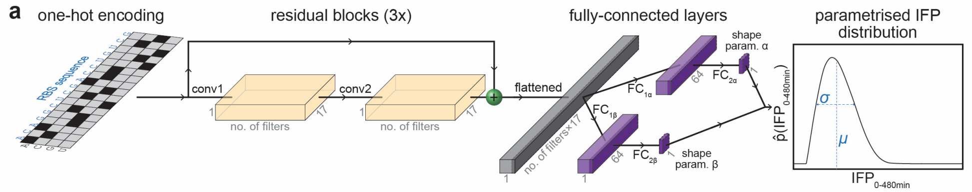
- `uASPIre`: new approach for sequencing-based phenotype recording for studying RBS activity in bacteria.
- Generates datasets of 100,000s of RBSs with activity phenotype
- Machine learning task: Can we use this data to make accurate predictions for *any possible* given RBS sequence?



Ribosome binding site activity prediction

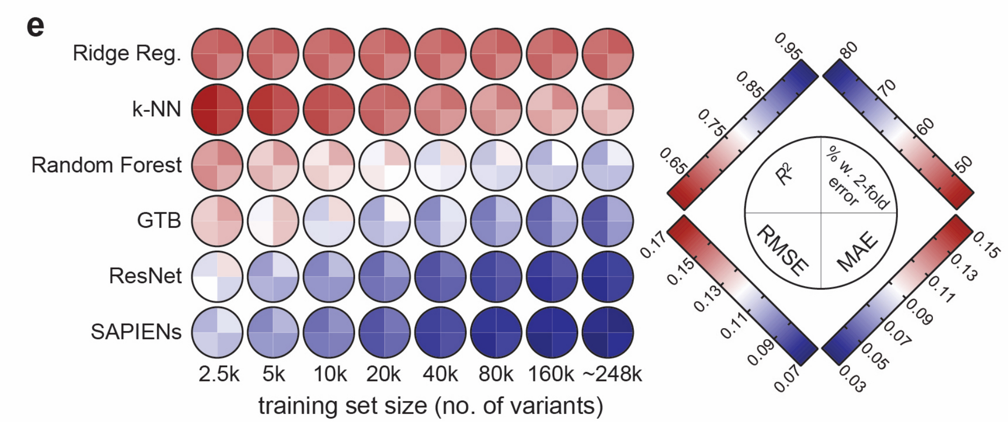
Methodological approach

- We developed a neural network to predict RBS activity from sequence: SAPIENs: Sequence-Activity Prediction In Ensemble of Networks



Ribosome binding site activity prediction

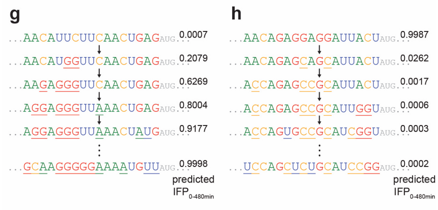
- Deep learning (SAPIENs) enables highly accurate sequence-function mapping



Ribosome binding site activity prediction

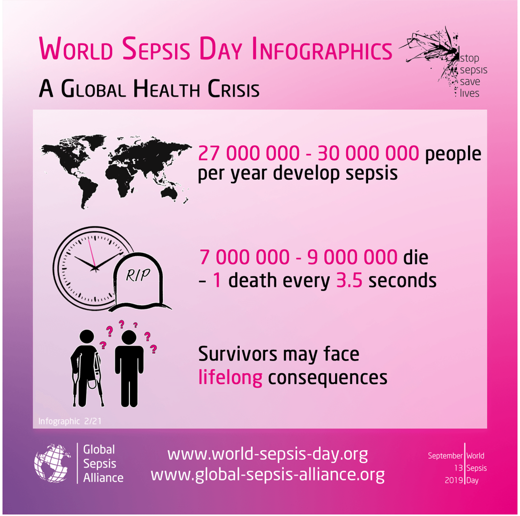
Current and future challenges

- Interpretation of SAPIENs predictions
- Design of RBS sequences using SAPIENs
- Integration of cellular context into SAPIENs
- Generalization to other gene regulatory elements



Machine learning in medicine

What is Sepsis?



Predicting Sepsis

Sepsis-3 definition (Singer et al., 2016)

- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection.

Relevance of early recognition

- Bacterial species identification in blood still takes 24h-48h (Osthoff et al., 2017).
- Each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

Predicting Sepsis

Sepsis-3 definition (Singer et al., 2016)

- Sepsis is a life-threatening organ dysfunction, caused by a dysregulated host response to infection.

Relevance of early recognition

- Bacterial species identification in blood still takes 24h-48h (Osthoff et al., 2017).
- Each hour of delayed effective antibiotic treatment increases mortality (Ferrer et al., 2014).

→ **Detecting and treating sepsis earlier** is of highest clinical interest.

Hectic fever, at its inception, is difficult to recognize but easy to treat; left unattended, it becomes easy to recognize and difficult to treat.

(Niccolò Machiavelli, Il Principe)

Predicting clinical outcomes in intensive care units

Input: patients' ICU data

- temperature
- heart rate
- blood pressure
- respiratory rate
- O_2 saturation

Output: sepsis prediction

- onset
- septic shock
- mortality

Predicting sepsis through time series classification

What is the state of the art in sepsis detection using ML?

Ref	Dataset	Label	Method	3h AU-ROC /-PR	Prev (%)
Futoma et al., 2017	Duke	Sepsis-2 'related'	MGP-RNN	0.96 / 0.87	21.4
Calvert et al., 2016	MIMIC-2	ICD-9 + 5h SIRS	InSight	0.92	11.4
Kam et al., 2017	MIMIC-2	ICD-9 + 5h SIRS	LSTM	0.93	6.6
Desautels et al., 2016	MIMIC-3	Sepsis-3	InSight eval	0.76 / 0.29	11.3

Predicting sepsis through time series classification

What is the state of the art in sepsis detection using ML?

Ref	Dataset	Label	Method	3h AU-ROC /-PR	Prev (%)
Futoma et al., 2017	Duke	Sepsis-2 'related'	MGP-RNN	0.96 / 0.87	21.4
Calvert et al., 2016	MIMIC-2	ICD-9 + 5h SIRS	InSight	0.92	11.4
Kam et al., 2017	MIMIC-2	ICD-9 + 5h SIRS	LSTM	0.93	6.6
Desautels et al., 2016	MIMIC-3	Sepsis-3	InSight eval	0.76 / 0.29	11.3

Critical care

BMJ Open Respiratory Research

Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial

David W Shimabukuro,[1] Christopher W Barton,[2] Mitchell D Feldman,[3] Samson J Mataraso,[4,5] Ritankar Das[6]

Shimabukuro *et al.*. BMJ Open Resp Res 2017;4:e000234.

Predicting sepsis through time series classification

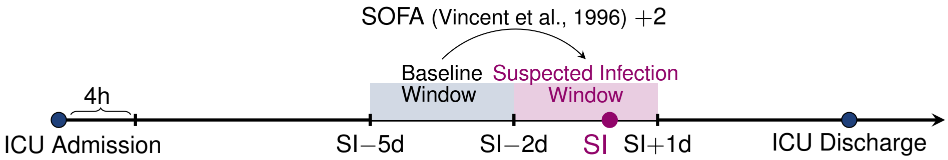
What is the state of the art in sepsis detection using ML?

- Johnson et al. (2018) showed that various sepsis definitions lead to different cohorts
- Low comparability due to heterogeneous phenotype definitions and implementations:
  - Several authors use ICD-9 billing code as sepsis label, without exact time of sepsis onset (e.g. Calvert et al., 2016, Kam et al., 2017)
  - Even for Sepsis-3 on MIMIC-III, the number of sepsis cases differs between studies:
    - 5,784 (Johnson et al., 2018),
    - 1,840 (Desautels et al., 2016),
    - 17,898 (Raghu et al. 2017)

Predicting sepsis through time series classification

Sepsis-3 definition

- **Case**
  - SI: suspicion of infection
  - SOFA: Sepsis-related organ failure assessment score



- **Control**
  - Only SI, or only SOFA score increase, or neither of them

Predicting sepsis through time series classification

Challenges

- **Comparability**
  - Heterogeneous label definitions (some insufficient for early detection task)
  - Heterogeneous label extraction (even on the same data with identical definition)
- **Reproducibility**
  - Unavailability of code for label extraction
- **Circularity**
  - Same observations used for prediction and definition of sepsis
- **Evaluation**
  - Time horizon analysis: which point in time to use for controls?
  - Few studies report precision / recall despite considerable class imbalance

Systematic review: Moor*, Rieck* et al., Frontiers in Medicine 2021

https://doi.org/10.3389/fmed.2021.607952

Early onset prediction based on Sepsis-3 definition

Moor et al., MLHC 2019

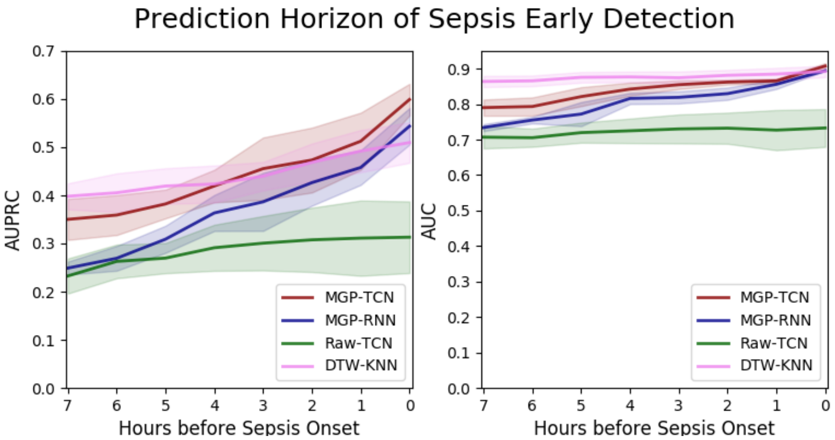
1. Determine temporally resolved Sepsis-3 labels on MIMIC-III
2. Imputation and regularization of measurements with Multi-Task Gaussian Processes
3. Classification with a Temporal Convolutional Network (MGP-TCN).
4. Classification with a Data Mining approach: Dynamic Time Warping k-nearest Neighbor (DTW-KNN) ensemble.

MIMIC-III dataset (after filtering)

Variable	Sepsis Cases	Controls
n	570	5,618
Female	236 (41.4%)	2,548 (45.4%)
Male	334 (58.6%)	3,070 (54.6%)
Mean time to sepsis onset in ICU (median)	16.7 h (11.8 h)	—
Age ($\mu \pm \sigma$)	67.2 ± 15.3	64.2 ± 17.3

Results

Early onset prediction on MIMIC-III (Moor et al., MLHC 2019)



Summary

Lessons we have learned

- Inherent challenges regarding comparability, reproducibility, circularity and proper evaluation
- Imputation scheme matters $\rightarrow$ methods for working on irregularly sampled time series are promising (Horn et al., ICML 2020)
- Deep learning architecture matters
- Classic baseline is the best early predictor $\rightarrow$ never miss to have a classic baseline

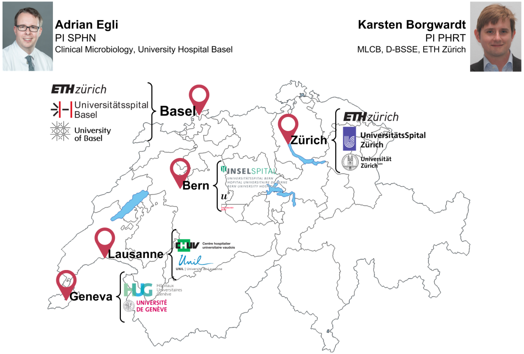
Current work: Personalized Swiss Sepsis Study

Goal

- Predict whether a patient will develop sepsis during ICU stay
  - Phase I: using clinical routine data
  - Phase II: using omics profiles

Current state

- Phase I: 10.000 health records collected across Switzerland
- Phase II: started recently



Moor et al., 2019, Moor et al., 2021

Current work: Wheat yield prediction

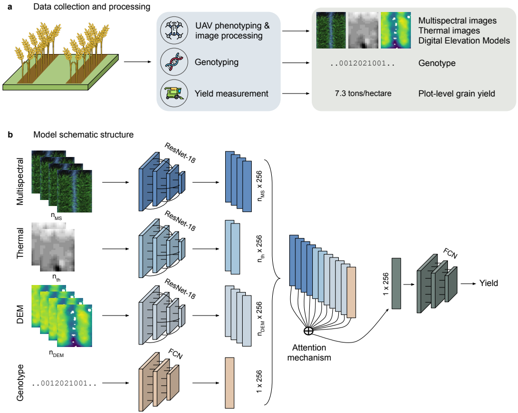
Goal

- Select wheat lines that provide high yield across environments

Current state

- Deep learning can drastically improve yield prediction when combining genotype and drone images

(Pearson's correlation 0.373 vs 0.026 linear model)



Togninalli et al., under preparation 2021

Machine learning in systems biology

Outlook

1. Biomarker discovery: predicting the phenotype of a system
2. Data integration: combining local and (massive) public datasets, different data types, accounting for confounding
3. Machine learning on structured data will be key to solving these problems

Future challenge: enormous data growth

- Sample size: reaching new magnitudes, from cell biology to medicine
- Time: more and longer longitudinal data
- Depth: multi-omics, or from lower- to higher-phenotypic level

Thank you

- Collaborators: Jeschek and Benenson labs at D-BSSE, PSSS consortium
- Sponsors: ERC-backup Scheme of Swiss National Science Foundation, Krupp-Stiftung, European Union (MSCA), SPHN/PHRT, SNSF, Botnar Foundation

References I

- J. Futoma, *et al.*, *arXiv preprint arXiv:1706.04152* (2017).
- J. S. Calvert, *et al.*, *Computers in Biology and Medicine* **74**, 69 (2016).
- H. J. Kam, H. Y. Kim, *Computers in biology and medicine* **89**, 248 (2017).
- T. Desautels, *et al.*, *JMIR Medical Informatics* **4**, e28 (2016).
- A. E. Johnson, *et al.*, *Critical care medicine* **46**, 494 (2018).
- A. Raghu, *et al.*, *arXiv preprint arXiv:1711.09602* (2017).
- K. Borgwardt, *et al.*, *Bioinformatics* **21**, i47 (2005).
- K. M. Borgwardt, H.-P. Kriegel, *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA* (IEEE Computer Society, 2005), pp. 74–81.
- K. Borgwardt, *et al.*, *Foundations and Trends® in Machine Learning* **13**, 531 (2020).
- E. Callaway, *Nature News* **588**, 203 (2020).
- R. Ferrer, *et al.*, *Critical Care Medicine* **42**, 1749 (2014).

References II

- S. Höllerer, *et al.*, *Nature Communications* **11**, 3551 (2020).
- M. Horn, *et al.*, *International Conference on Machine Learning* (PMLR, 2020), pp. 4353–4363.
- S. L. Hyland, *et al.*, *Nature Medicine* **26**, 364 (2020).
- T. N. Kipf, M. Welling, *arXiv preprint arXiv:1609.02907* (2016).
- T. A. Manolio, *et al.*, *Nature* **461**, 747 (2009).
- M. Moor, *et al.*, *Machine Learning for Healthcare Conference* (2019), pp. 2–26.
- M. Moor, *et al.*, *Frontiers in Medicine* **8** (2021).
- L. O'Bray, *et al.*, *arXiv:2106.01098 [cs, stat]* (2021).
- M. Osthoff, *et al.*, *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **23**, 78 (2017).
- N. Pržulj, *et al.*, *Bioinformatics* **22**, 974 (2006).
- C. W. Seymour, *et al.*, *JAMA* **315**, 762 (2016).

References III

- N. Shervashidze, K. M. Borgwardt, *Advances in Neural Information Processing Systems 22:*, Y. Bengio, *et al.*, eds. (Curran Associates, Inc., Vancouver, British Columbia, Canada, 2009), pp. 1660–1668.
- N. Shervashidze, *et al.*, *Journal of Machine Learning Research* **12**, 2539 (2011).
- M. Singer, *et al.*, *JAMA* **315**, 801 (2016).
- J. M. Stokes, *et al.*, *Cell* **180**, 688 (2020).
- M. Sugiyama, *et al.*, *Bioinformatics* **34**, 530 (2018).
- J.-L. Vincent, *et al.*, *Intensive Care Medicine* **22**, 707 (1996).