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
Further central topics

- **Chemoinformatics**: predicting function based on molecular structure
Further central topics

- **Chemoinformatics**: predicting function based on molecular structure
- **Medicine**: predicting disease diagnosis, progression, therapy outcome
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
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**
Obstacles for machine learning in the life sciences

1. Not enough observations
Obstacles for machine learning in the life sciences

1. Not enough observations
2. Uncertainty and difficulty in phenotyping
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
Recently big progress

- **Protein structure prediction**
  
  *Nature* - 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**

  *Cell* - 19 February 2020

  A Deep Learning Approach to Antibiotic Discovery

  Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Tim S. Jaakkola, Regina Barzilay, James J. Collins

  DOI: https://doi.org/10.1016/j.cell.2020.01.021
Recently big progress

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

- Molecular function prediction
  - A Deep Learning Approach to Antibiotic Discovery

Both use machine learning on graphs
Machine learning on graphs
Graphs are the data structure to represent systems, networks and structures.
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)
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)

Given labeled graphs $G_1$ and $G_2$...

Shervashidze et al., 2011
Machine learning on graphs

Fundamental question: How similar are two 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)

Filtration Sequence

\[ G_1 \quad w_1 = 10 \]
\[ G_2 \quad w_2 = 20 \]
\[ G_3 \quad w_3 = 50 \]
\[ G_4 \quad w_4 = 80 \]
\[ G_5 = G \quad w_5 = 90 \]

Filtration Curves

Node label count

Edge weight \( w_i \)
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 $F_G$
   - (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 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 \( F_G = (G_1, \ldots, G_m) \).
  - and a graph descriptor function \( f : G \rightarrow \mathbb{R}^d \)

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

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

(1)

- **where**
  - \( m \) indexes the number of edge weight thresholds in \( F_G \), 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Abs. distance to accuracy of best method</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSM</td>
<td>0.00</td>
</tr>
<tr>
<td>HGK-SP</td>
<td>0.05</td>
</tr>
<tr>
<td>HGK-WL</td>
<td>0.15</td>
</tr>
<tr>
<td>MLG</td>
<td>0.10</td>
</tr>
<tr>
<td>WL</td>
<td>0.20</td>
</tr>
<tr>
<td>WL-OA</td>
<td>0.25</td>
</tr>
<tr>
<td>GNN</td>
<td>0.15</td>
</tr>
<tr>
<td>FC</td>
<td>0.20</td>
</tr>
</tbody>
</table>

---

ETH Zürich  
Karsten Borgwardt (@kmborgwardt)

Turkish Science Academy  | June 23, 2021 | 17 / 44
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 et al., 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)

Inherently related to learning on sequences, time series and images - which also have manifold (potential) applications in the life sciences
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
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
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
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?

**WORLD SEPSIS DAY INFOGRAPHICS**

**A GLOBAL HEALTH CRISIS**

- 27,000,000 - 30,000,000 people per year develop sepsis
- 7,000,000 - 9,000,000 die
  - 1 death every 3.5 seconds

Survivors may face lifelong consequences

[World Sepsis Day Infographics](https://www.world-sepsis-day.org)
[Global Sepsis Alliance](https://www.global-sepsis-alliance.org)
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
- O₂ 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?

<table>
<thead>
<tr>
<th>Ref</th>
<th>Dataset</th>
<th>Label</th>
<th>Method</th>
<th>3h AU-ROC /-PR</th>
<th>Prev (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Futoma et al., 2017</td>
<td>Duke</td>
<td>Sepsis-2 'related'</td>
<td>MGP-RNN</td>
<td>0.96 / 0.87</td>
<td>21.4</td>
</tr>
<tr>
<td>Calvert et al., 2016</td>
<td>MIMIC-2</td>
<td>ICD-9 + 5h SIRS</td>
<td>InSight</td>
<td>0.92</td>
<td>11.4</td>
</tr>
<tr>
<td>Kam et al., 2017</td>
<td>MIMIC-2</td>
<td>ICD-9 + 5h SIRS</td>
<td>LSTM</td>
<td>0.93</td>
<td>6.6</td>
</tr>
<tr>
<td>Desautels et al., 2016</td>
<td>MIMIC-3</td>
<td>Sepsis-3</td>
<td>InSight eval</td>
<td>0.76 / 0.29</td>
<td>11.3</td>
</tr>
</tbody>
</table>
## Predicting sepsis through time series classification

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

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

---

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

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

Prediction Horizon of Sepsis Early Detection
Summary

Lessons we have learned

- Inherent challenges regarding comparability, reproducibility, circularity and proper evaluation
- Imputation scheme matters → 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 → 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
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)
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

References II


