TUDataset: A collection of benchmark datasets for learning with graphs

Christopher Morris[1] **Nils M. Kriege**[2] Franka Bause[3]
Kristian Kersting[4] Petra Mutzel[5] Marion Neumann[6]

[1]Polytechnique Montréal

[2]University of Vienna

[3]TU Dortmund University

[4]TU Darmstadt

[5]University of Bonn

[6]Washington University in St. Louis

July 17, 2020

Graph Classification and Regression

Setting

Learn function $f : \mathcal{G} \to \mathcal{Y}$ predicting a property of a graph.

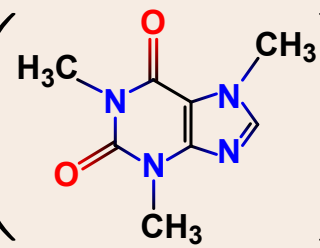
Graph Classification and Regression

Setting

Learn function $f : \mathcal{G} \to \mathcal{Y}$ predicting a property of a graph.

Example

- $\mathcal{G}$ – set of molecular graphs
- $\mathcal{Y} = \{\texttt{toxic}, \texttt{non-toxic}\}$

$$f\left(\text{  }\right) = ?$$

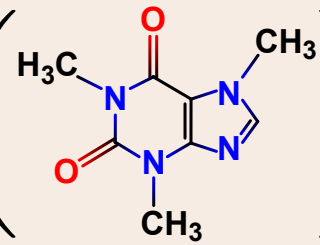
Graph Classification and Regression

Setting

Learn function $f : \mathcal{G} \to \mathcal{Y}$ predicting a property of a graph.

Example

- $\mathcal{G}$ – set of molecular graphs
- $\mathcal{Y} = \{\texttt{toxic}, \texttt{non-toxic}\}$

$$f\left(\text{  }\right) = ?$$

Approaches:

- Simple vector representations, e.g. chemical fingerprints (1973)
- Graph kernels (2002)
- Graph neural networks (2016)

Reliable Experimental Evaluations

Problems

- Graph models non-standardized
- Experimental setup non-standardized

$\Rightarrow$ Results from different publications not comparable

Reliable Experimental Evaluations

Problems

- Graph models non-standardized
- Experimental setup non-standardized

$\Rightarrow$ Results from different publications not comparable

Common weak points

- Missing comparison with baseline methods
- Used benchmark data sets are
  - too small
  - too easy to solve
  - insufficiently diverse

TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

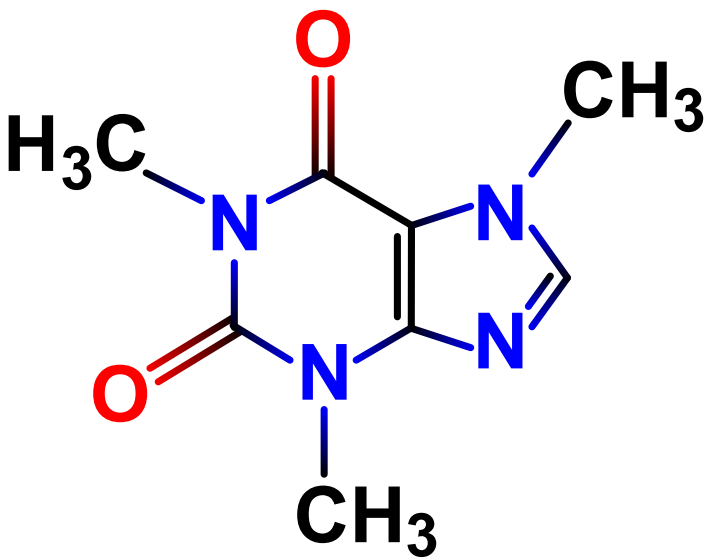
- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic

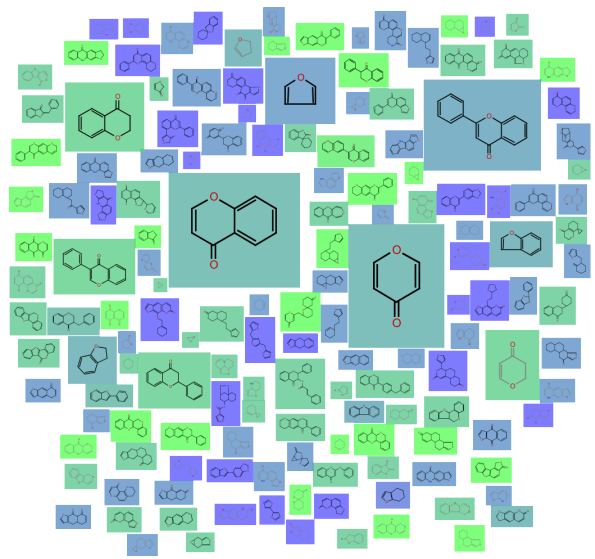
TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic



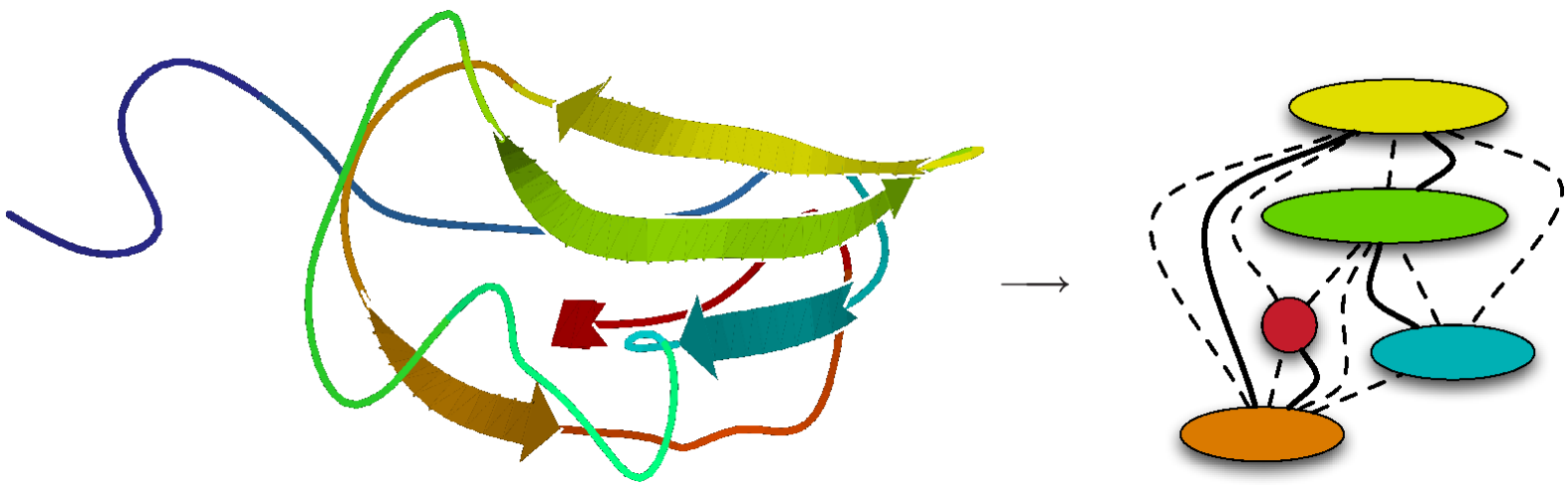


TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic



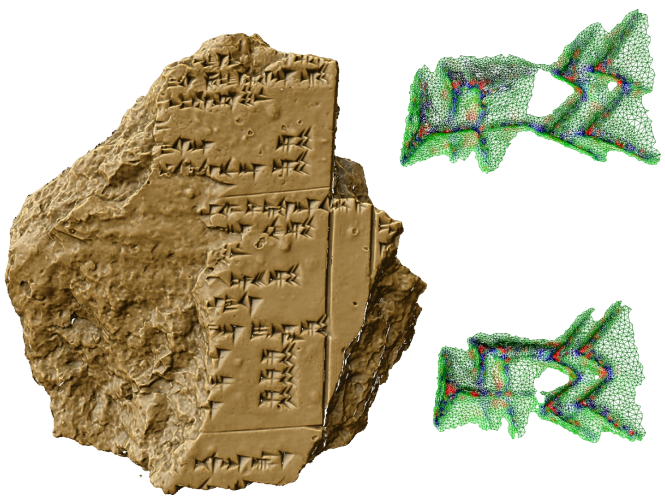
[Borgwardt et al., 2005; Vishwanathan et al., 2010]

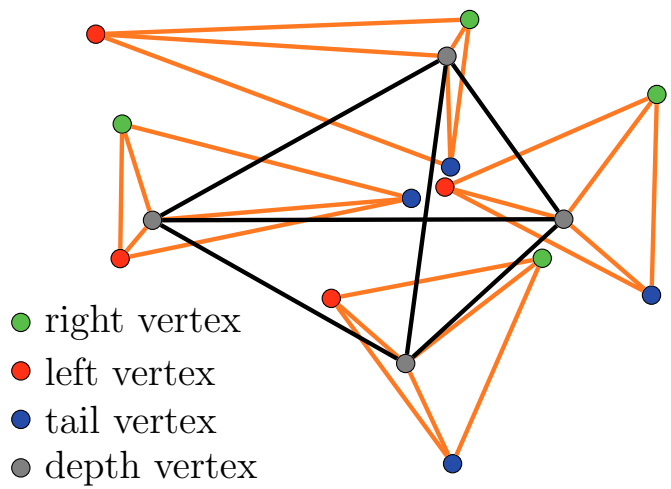
TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic





[Kriege et al., 2018]

TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic

TUDataset Benchmark Collection

(1) Datasets

Over 120 graph datasets of various sizes and domains:

- Small molecules
- Bioinformatics
- Computer vision
- Social networks
- Synthetic

TUDataset Benchmark Collection

(2) Baseline methods

- Shortest-path kernel [Borgwardt, Kriegel,2005]
- Graphlet [Shervashidze et al., 2009]
- Weisfeiler-Lehman subtree kernel [Shervashidze et al., 2011]
- Weisfeiler-Lehman optimal assignment kernel [Kriege et al., 2016]
- GNN architectures of PyTorch Geometric [Fey, Lenssen, 2019]

TUDataset Benchmark Collection

(2) Baseline methods

- Shortest-path kernel [Borgwardt, Kriegel,2005]
- Graphlet [Shervashidze et al., 2009]
- Weisfeiler-Lehman subtree kernel [Shervashidze et al., 2011]
- Weisfeiler-Lehman optimal assignment kernel [Kriege et al., 2016]
- GNN architectures of PyTorch Geometric [Fey, Lenssen, 2019]

(3) Evaluation module

- 10-fold cross validation
- Hyperparameter optimization for each fold

Summary

- Collection of over 120 graph datasets
- Standard file format with data loaders
- Evaluation modules in Python
- Graph kernel baselines in C++ with Python bindings
- Accessible via graph learning frameworks

Summary

- Collection of over 120 graph datasets
- Standard file format with data loaders
- Evaluation modules in Python
- Graph kernel baselines in C++ with Python bindings
- Accessible via graph learning frameworks

`http://graphlearning.io`

Summary

- Collection of over 120 graph datasets
- Standard file format with data loaders
- Evaluation modules in Python
- Graph kernel baselines in C++ with Python bindings
- Accessible via graph learning frameworks

`http://graphlearning.io`

Thank You!

- Many thanks to all contributors of datasets
- Your involvement is greatly appreciated!