group of machine gmum learning research

Learning to SMILE(s)

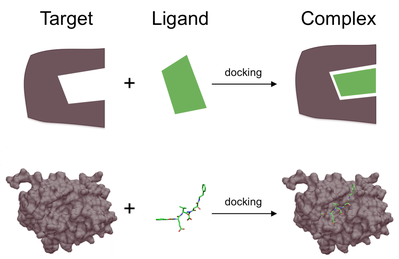
Stanislaw Jastrzebski, Damian Lesniak, Wojciech Marian Czarnecki
Jagiellonian University, Cracow, Poland

Drug discovery + Deep Learning

Drug development can take **10 years** and consume over **2B$**. Computer methods are **fundamental** in modern drug discovery, but **DL has had limited success**. Why?

Virtual Screening

Virtual Screening is a vital part of computer aided drug discovery. It is used to filter drug candidates before doing actual *very expensive and long* tests in laboratory.



- Here binary classification: **Does the molecule bind to the target?**
- Input is a small graph.

Target is usually a protein found in human body, for instance responsible for transfering electrical signal to cell.

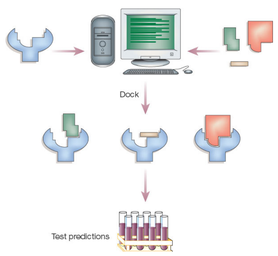
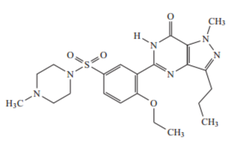


Fig. 2: Docked compound

Molecules

Input is a small molecule (usually ≤ 100 atoms).
Most methods can't take graph as input, standard approach is to convert it to a constant sized vector: a **"fingerprint"**.



keys					
fingerprint	1	0	1	1	1

Hand-crafted representations are often *bad*.

Text classifiers on molecules

Molecules are normally stored in databases in **SMILES** format - a string of characters forming the compound. Example:

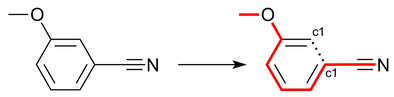
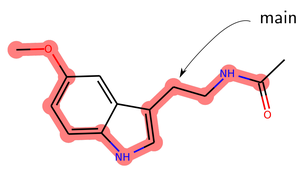


Fig. 4: `N(c1)ccc1N`

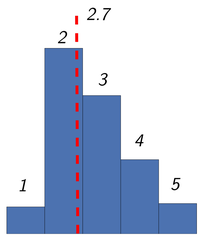
This is a raw representation, it shouldn't perform better than hand-crafted ones... or **maybe**?

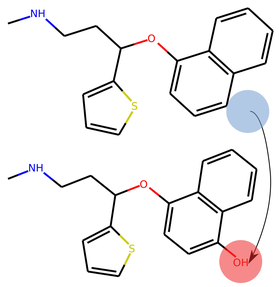
Turn out there are **analogies** to sentiment classification:



1. We treat main walk as the **main clause** and branches as subclauses.

2. Average distance to the main branch of the tree is under 2! Almost all molecules are **tree-like**.





3. *"I like the movie"* vs *"I **don't** like the movie"*. Similarly as in sentiment classification, activity is very sensitive to local changes.

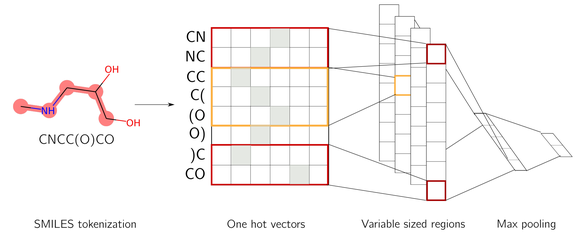
Experiments

We evaluate against state of the art **substructural** fingerprints on 5 fairly small binary datasets. Models are selected to cover popular choices in traditional virtual screening approaches and NLP.

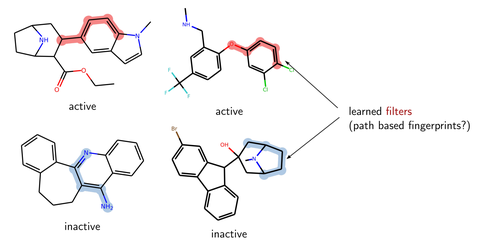
- **Traditional models**: SVM, Random Forest, Naive Bayes.
- **NLP models**: RNN (GRU), one hot CNN, Recurrent Neural Language Model (RNNLM).

CNN on molecules

Best model: one hot CNN.

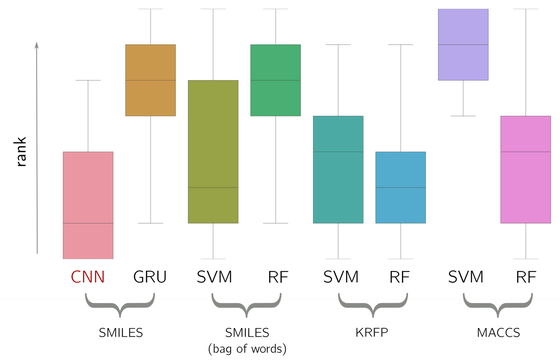


It can be interpreted as **automatically learning representation**.



Results

Models ranks by log-loss (RNNLM and NB were omitted for clarity).



Conclusions & future directions

- This is **work in progress**. Stronger hand-crafted fingerprints and more models have to be tested.
- Very promising direction for **semi-supervised learning**

Acknowledgments: First author was supported by Grant No. DI 2014/016644 from Ministry of Science and Higher Education, Poland.