Insight
Centre for Data Analytics

NUI Galway
OÉ Gaillimh

FUJITSU
shaping tomorrow with you

Using Drug Similarities for Discovery of Possible Adverse Reactions

Emir Muñoz

Fujitsu (Ireland) Limited, *Researcher*

Insight Centre for Data Analytics at NUI Galway, *PhD Student*

Joint work with Vít Nováček and Pierre-Yves Vandenbussche

November 15th, 2016 AMIA, Chicago, US

Disclosure

I receive funding from:

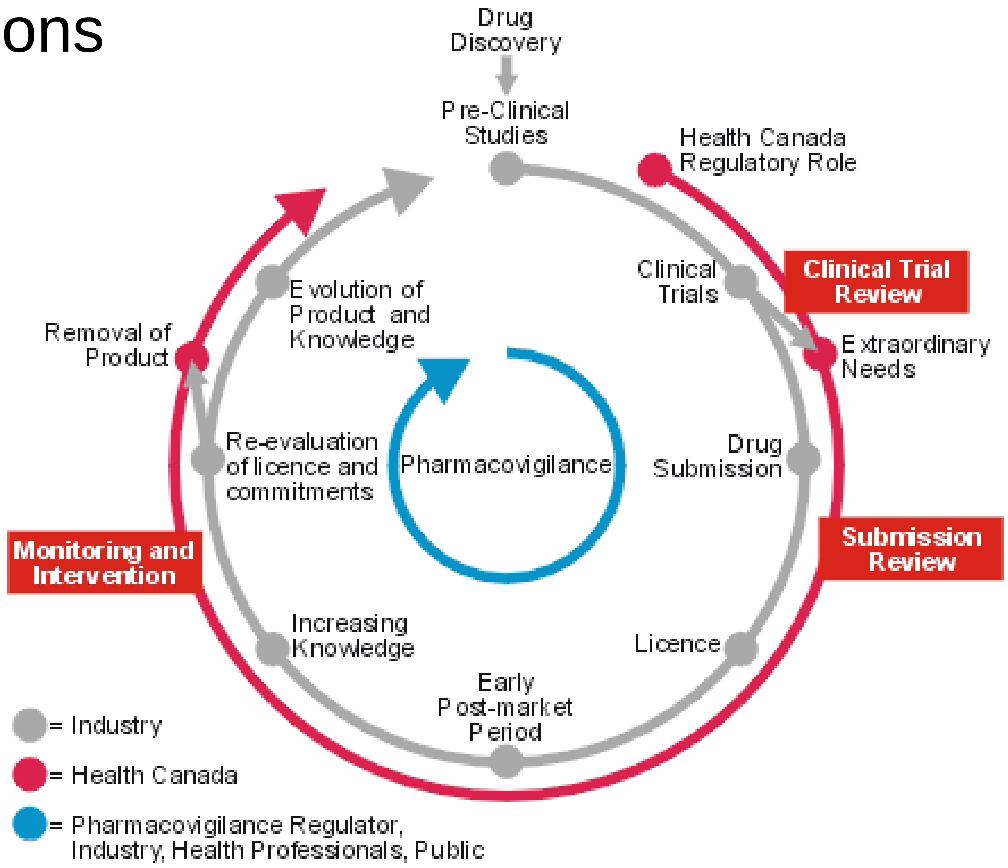
- Fujitsu Laboratories, Japan
- Insight Centre for Data Analytics, NUI Galway, Ireland

Learning Objectives

- Improve Adverse Drug Events detection using propagation of known side effects between similar drugs.
- Formulate an extensible approach for Adverse Drug Events detection using linked open data sources.

Introduction (1/2)

- Drug development is an expensive process
- Adverse drug reactions (ADR) account for 42% of hospital admissions
- Most ADRs are reported after commercialization



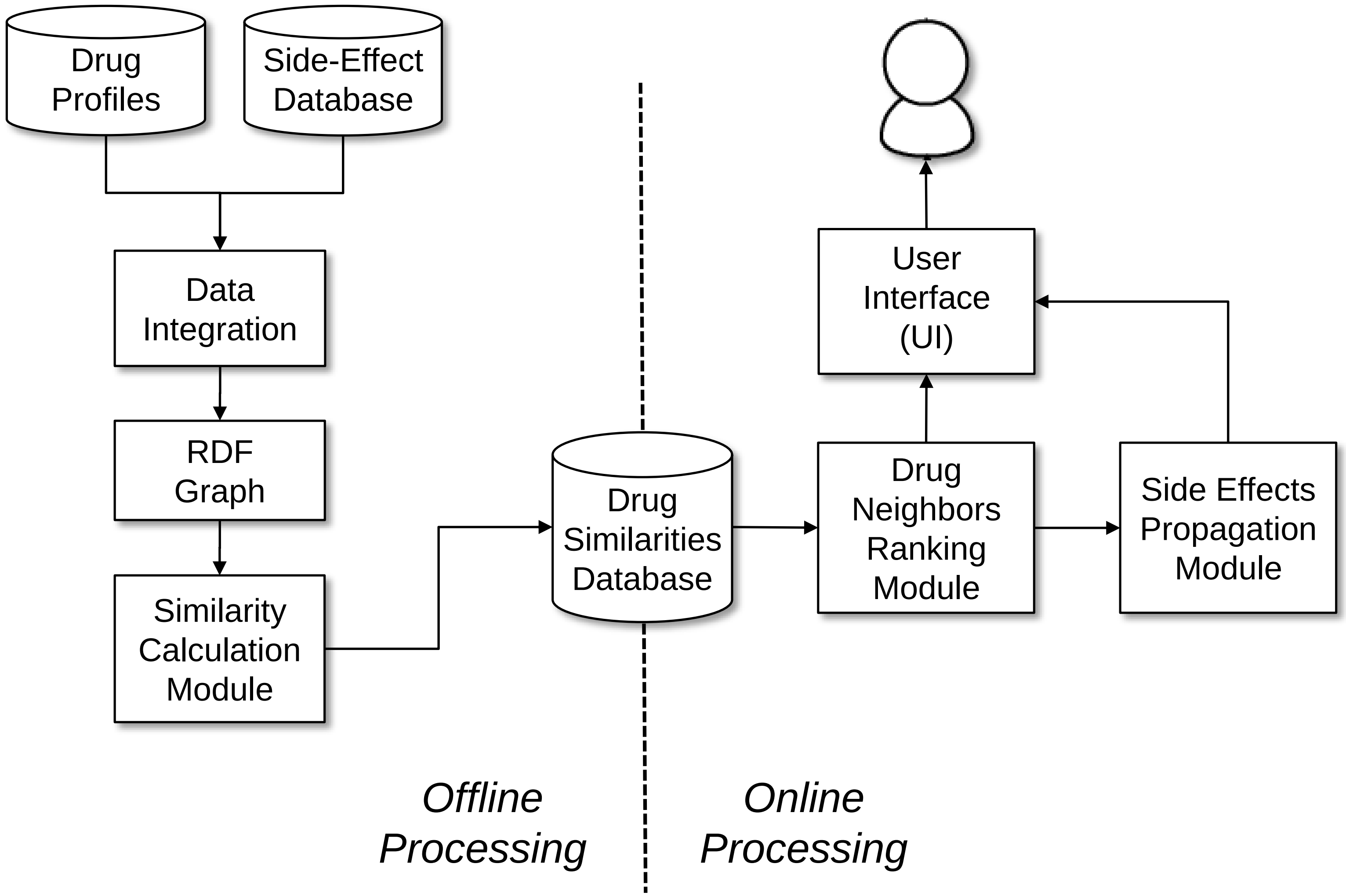
Health Canada: http://www.hc-sc.gc.ca/dhp-mps/homologation-licensing/model/life-cycle-vie-eng.php

Introduction (2/2)

- Problem: Discover relations between drugs and ADRs
- Assumption: Similar drugs share a set of ADRs

- ADRs can be propagated from one drug to its most similar neighbors
- SoA approaches represent drugs using feature vector representations from *isolated sources*:
  - Enzyme, Pathway, Target, Transporter, Indication, and Substructure
- We believe that knowledge integrated from different data sources can provide better results

System Overview

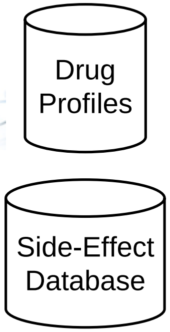

Drug Profiles
Side-Effect Database
Data Integration
RDF Graph
Similarity Calculation Module
Drug Similarities Database
User Interface (UI)
Drug Neighbors Ranking Module
Side Effects Propagation Module
Offline Processing
Online Processing

Methods (1/5)

Data sources

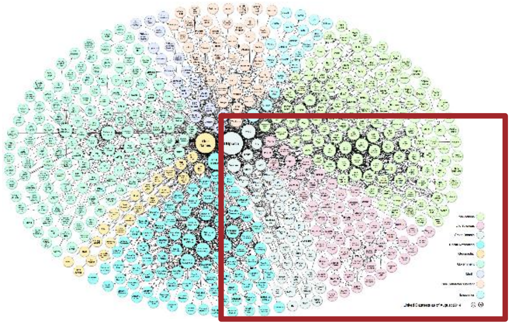
DRUGBANK
Open Data Drug & Drug Target Database

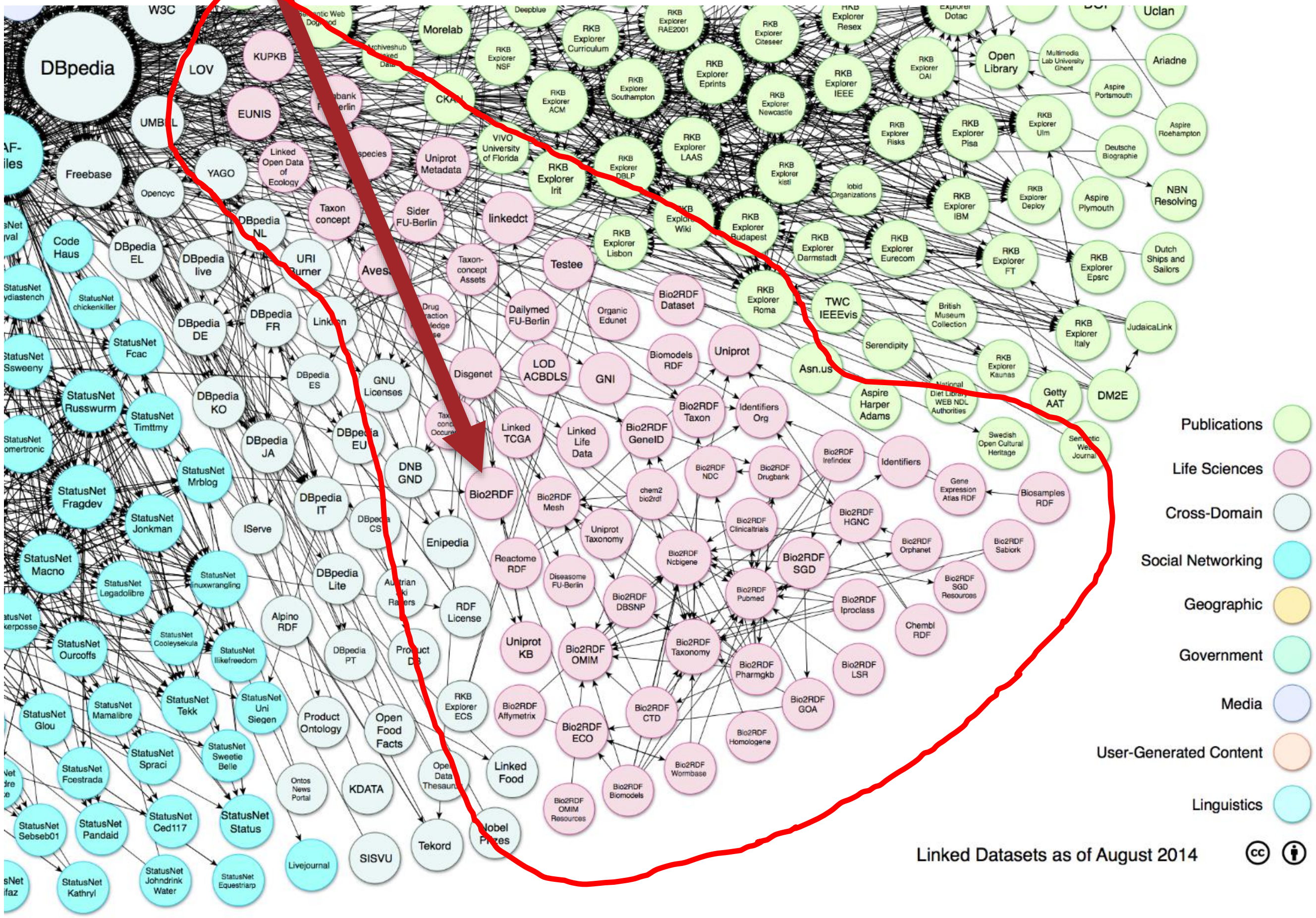
SIDER



BIO2RDF

http://download.openbiocloud.org/release/4/
(accessed in December 2015)





Linked Datasets as of August 2014

Methods (2/5)

Data processing

- DrugBank and SIDER are represented using RDF and queried using SPARQL
- 10.7 million unique statements represented as N-Quads (subject, predicate, object, graph)
- RDF triple store Apache Jena Fuseki2 (http://jena.apache.org/)
- Relevant statistics:
    - 731 approved small-molecule drugs
    - 4,652 side effects

Methods (3/5)

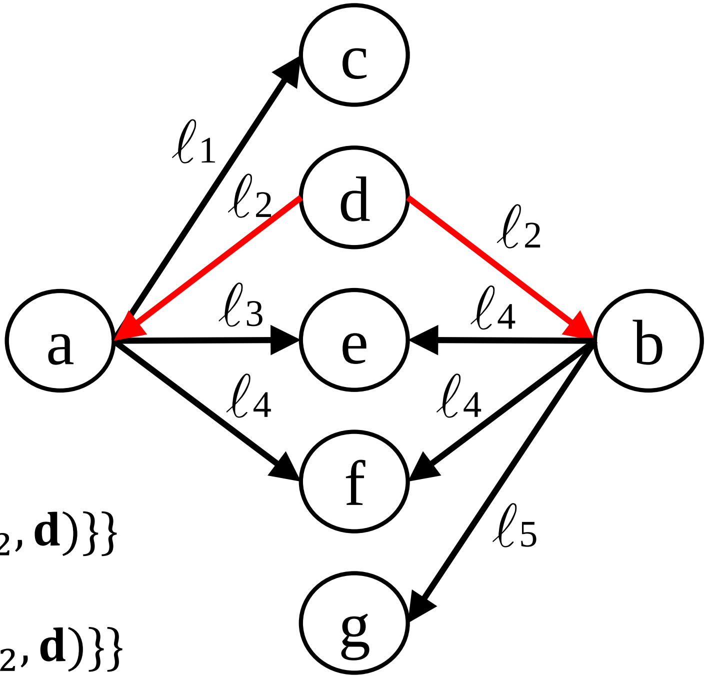
Measures

- *Resource features vector*
- Our features come from the graph structure
- We query the knowledge graph

using graph patterns as:

  - (?, ?, **X**) – incoming edges
  - (**X**, ?, ?) – outgoing edges

- Example:



$$A = Features_{LD}(\mathbf{a}) = \{\{(\ell_1, \mathbf{c}), (\ell_3, \mathbf{e}), (\ell_4, \mathbf{f})\}, \{(\ell_2, \mathbf{d})\}\}$$

$$B = Features_{LD}(\mathbf{b}) = \{\{(\ell_4, \mathbf{e}), (\ell_4, \mathbf{f}), (\ell_5, \mathbf{g})\}, \{(\ell_2, \mathbf{d})\}\}$$

Methods (4/5)

Measures

- With the feature vectors we can compute similarity
- Intuitively, the more features two nodes have in common, the more similar they are
- 3W-Jaccard similarity, defined as:

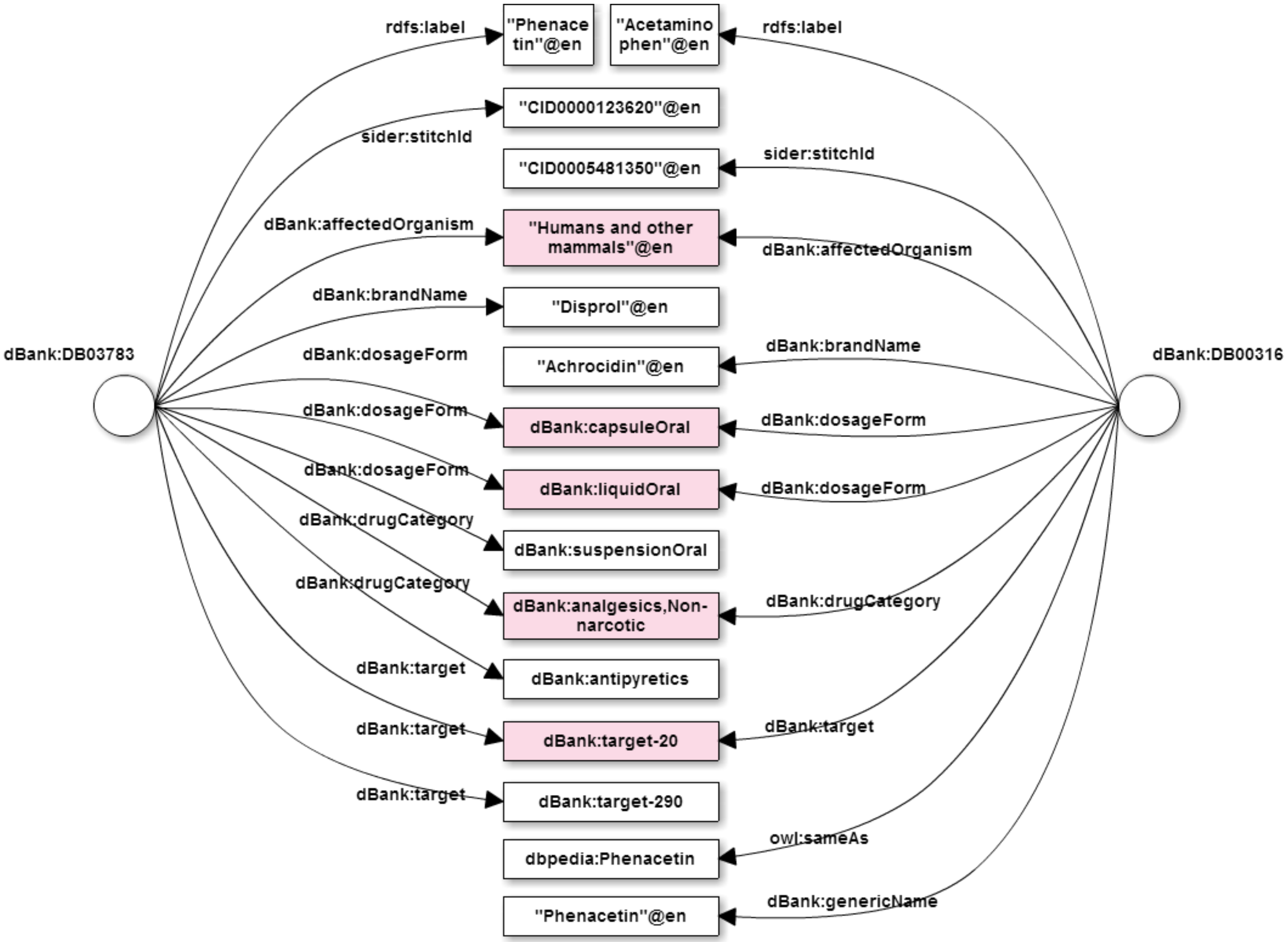
$$S_{3W-Jaccard}(a,b) = \frac{3x}{3x+y+z},$$

$$whith\ 0 \leq S_{3W-Jaccard}(a,b) \leq 1$$

$$x = |A \cap B|$$
$$y = |A - B|$$
$$z = |B - A|$$

  - Gives *high weight to common features*, and *lower weight to discriminating features*

Methods (5/5)



$$S_{3W-Jaccard}(dBank\text{:}\,DB03783, dBank\text{:}\,DB00316) = \frac{3\ \times 5}{3 \times 5 + 6 + 5} = 0.5769$$

Prediction of Side Effects (1/2)

Algorithm: Multi-label classification

- 1. Compute the features vector for each drug
- 2. Compute similarity between every pair of drugs
- 3. For each drug x_i extract the k neighborhood ($k = 50$)
  - 3.1. Filter neighborhood using a threshold $[0 - 1]$
  - 3.2. Propagate side effects in the k neighborhood to x_i

Let:

- $\mathbf{W}_{UL}$be the distance to x_i
- $\mathbf{L}_{UU}$ the sum of the distances
- $\mathbf{f}_L$ the vector of relative freq.

for a given side effect s in all neighbors

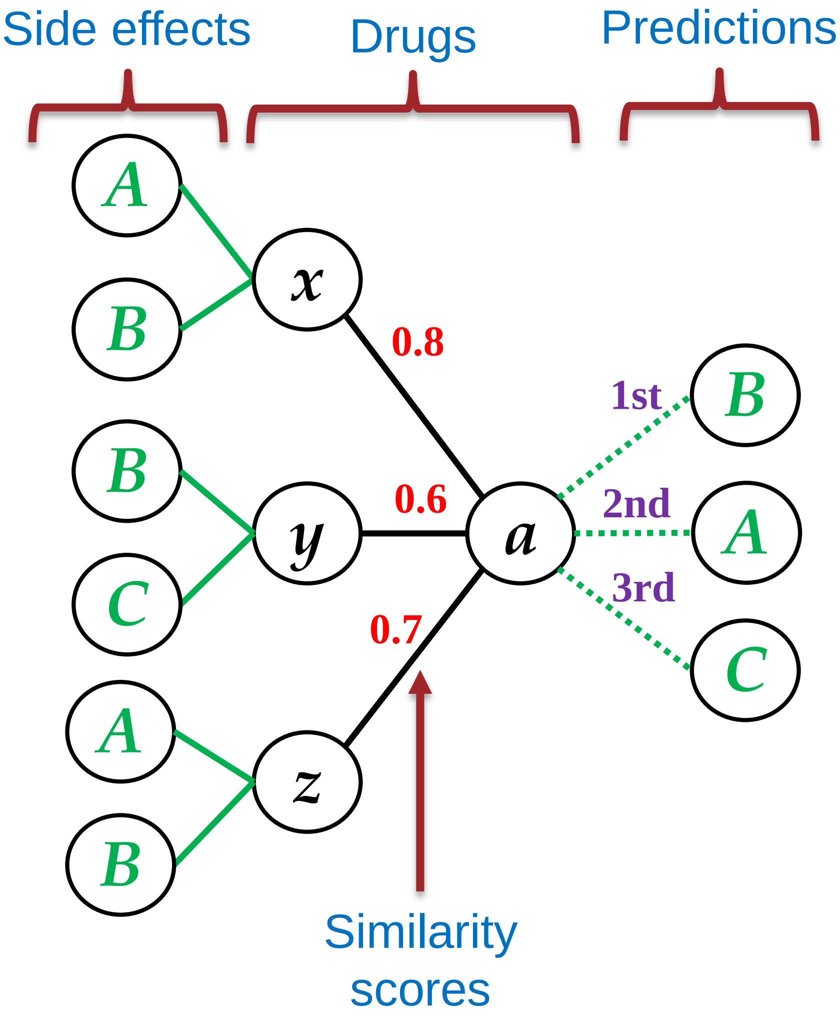
Side effect s propagation in drug x_i

$$s_{weight}(x_i) = \frac{1}{\mathbf{L}_{UU}} \mathbf{W}_{UL} \mathbf{f}_L$$

Prediction of Side Effects (2/2)

- Example: Predictions for drug $\boldsymbol{a}$

- $\mathbf{W}_{UL} = [0.8,\ 0.6,\ 0.7]$
- $\mathbf{L}_{UU} = 0.8 + 0.6 + 0.7 = 2.1$
- $\mathbf{f}_L^A = [1,\ 0,\ 1]^T$
- $\mathbf{f}_L^B = [1,\ 1,\ 1]^T$
- $\mathbf{f}_L^C = [0,\ 1,\ 0]^T$
- $A_{weight}(a) = \frac{1}{2.1} 1.5 = 0.7143$
- $B_{weight}(a) = \frac{1}{2.1} 2.1 = 1.0$
- $C_{weight}(a) = \frac{1}{2.1} 0.6 = 0.2857$



Results and Discussion (1/6)

Evaluation data set

- Leave-one-out cross validation

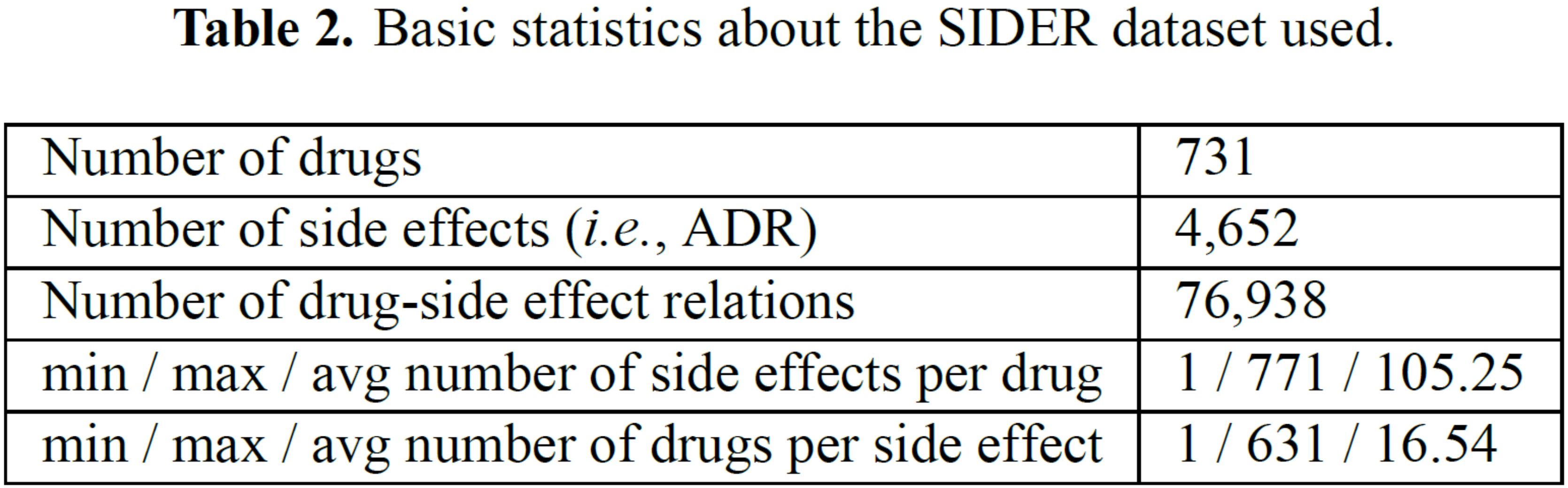


Table 2. Basic statistics about the SIDER dataset used.

Number of drugs	731
Number of side effects (*i.e.*, ADR)	4,652
Number of drug-side effect relations	76,938
min / max / avg number of side effects per drug	1 / 771 / 105.25
min / max / avg number of drugs per side effect	1 / 631 / 16.54

Results and Discussion (2/6)

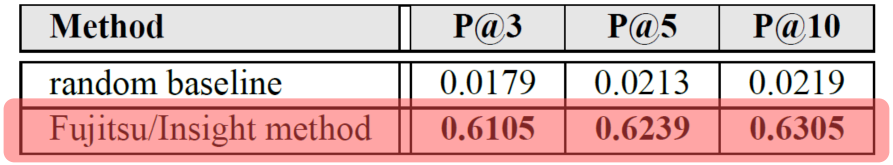
Evaluation Methodology

- Metrics for multi-label classification:
    - Precision
    - Recall
    - Accuracy
    - F1-score
    - Average precision
- We also focused on the ranking of the scores
    - Top1
    - Top5
    - P@3, P@5, P@10

Results and Discussion (3/6)

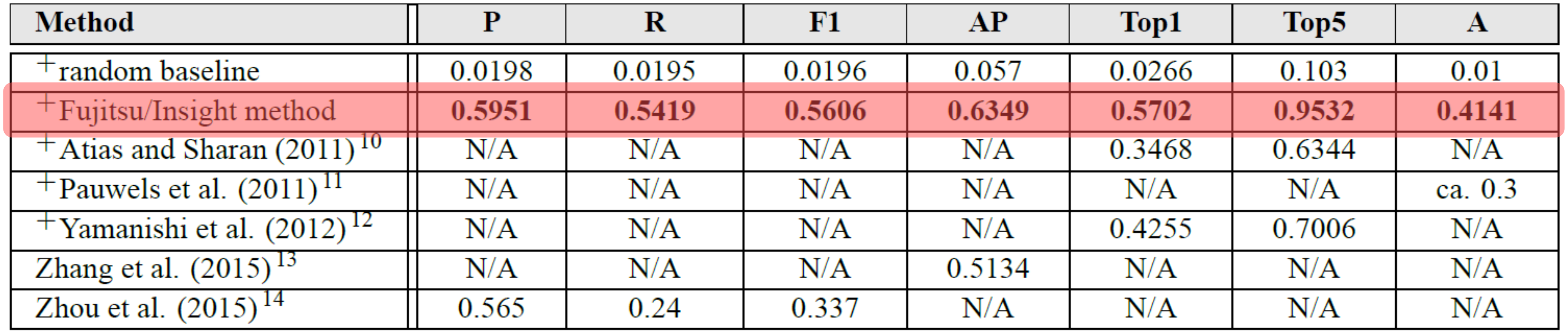
Results (threshold = 0.6)

Table 4. P@K results.



Method	**P@3**	**P@5**	**P@10**
random baseline	0.0179	0.0213	0.0219
Fujitsu/Insight method	**0.6105**	**0.6239**	**0.6305**

Table 3. Comparison of the results with related methods.



Method	**P**	**R**	**F1**	**AP**	**Top1**	**Top5**	**A**
+random baseline	0.0198	0.0195	0.0196	0.057	0.0266	0.103	0.01
+Fujitsu/Insight method	**0.5951**	**0.5419**	**0.5606**	**0.6349**	**0.5702**	**0.9532**	**0.4141**
+Atias and Sharan (2011) [10]	N/A	N/A	N/A	N/A	0.3468	0.6344	N/A
+Pauwels et al. (2011) [11]	N/A	N/A	N/A	N/A	N/A	N/A	ca. 0.3
+Yamanishi et al. (2012) [12]	N/A	N/A	N/A	N/A	0.4255	0.7006	N/A
Zhang et al. (2015) [13]	N/A	N/A	N/A	0.5134	N/A	N/A	N/A
Zhou et al. (2015) [14]	0.565	0.24	0.337	N/A	N/A	N/A	N/A

(Approximated comparison based on references' results)

Results and Discussion (4/6)

Results analysis

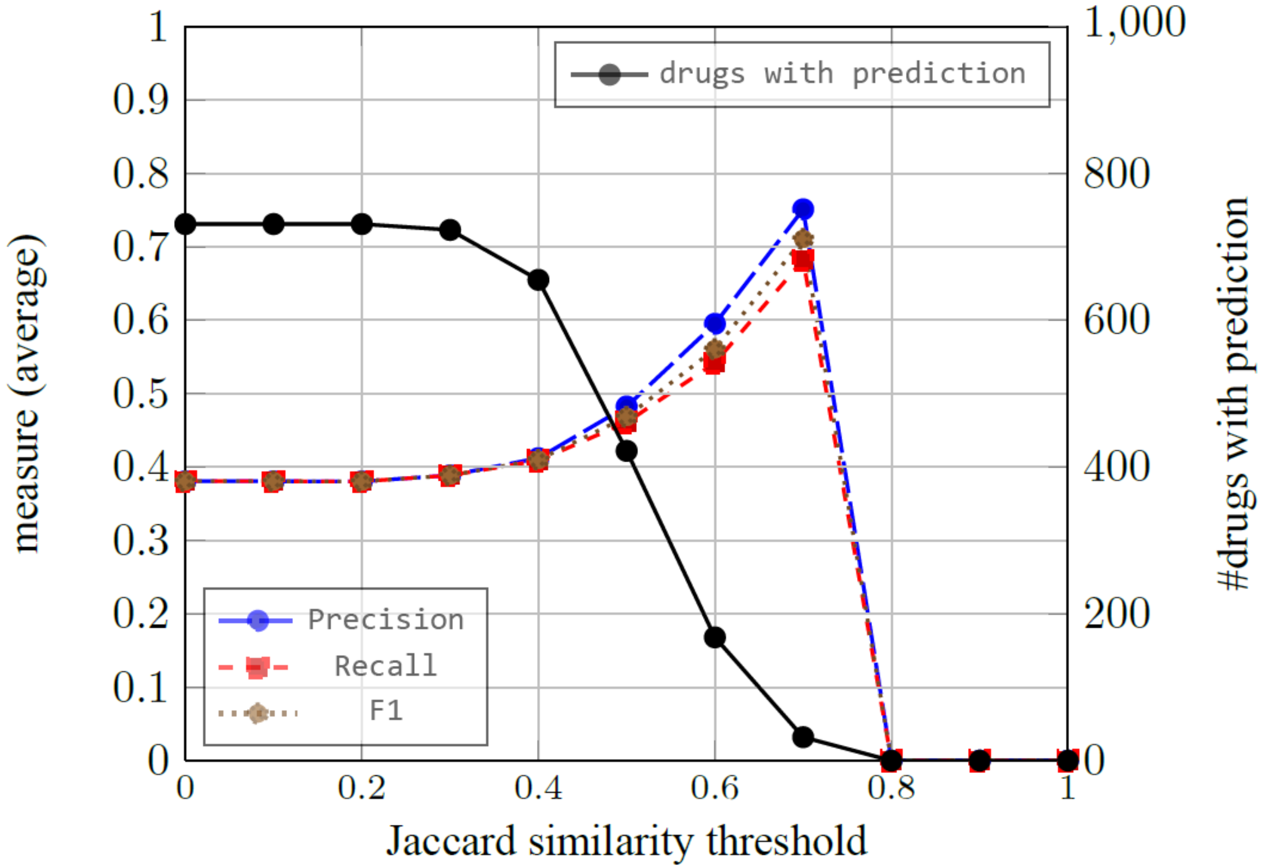


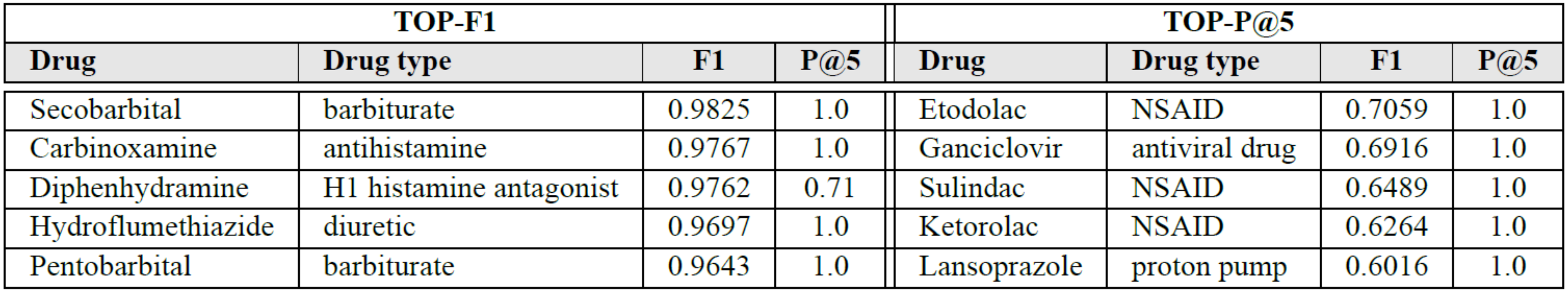
Figure 4. Plot of the results in relation to the similarity threshold.

Results and Discussion (5/6)

Examples of results

- We observed some frequent drug types among the best performing results: barbiturates, antihistamines and NSAIDs
  - This should be checked further in future works

Table 5. Examples of top-scoring drugs.



TOP-F1				TOP-P@5			
Drug	**Drug type**	**F1**	**P@5**	**Drug**	**Drug type**	**F1**	**P@5**
Secobarbital	barbiturate	0.9825	1.0	Etodolac	NSAID	0.7059	1.0
Carbinoxamine	antihistamine	0.9767	1.0	Ganciclovir	antiviral drug	0.6916	1.0
Diphenhydramine	H1 histamine antagonist	0.9762	0.71	Sulindac	NSAID	0.6489	1.0
Hydroflumethiazide	diuretic	0.9697	1.0	Ketorolac	NSAID	0.6264	1.0
Pentobarbital	barbiturate	0.9643	1.0	Lansoprazole	proton pump	0.6016	1.0

Results and Discussion (6/6)

Discussion

- Non-zero cut-offs decrease the number of predictions we can make
- Which delivers good results until the 0.6 cut-off
- Previous approaches treat the problem only as classification or only as ranking
  - We tried to mix both approaches and compare as much as we can
  - There is no clear gold-standard out there
  - SIDER seems to be the best option at the moment for sort of formal benchmarking
    - (We are working on a method using FDA reports and AEOLUS data set for complementary evaluation)

Conclusions and Future Work

Summarizing

- Similarity of drugs can be used to propagate adverse reactions
- Graph-based similarities show promising results

Next steps

- Propagation using graph regularization
  - Gaussian label propagation
- Inclusion of more drug- and disease- related Bio2RDF data sets in our knowledge graph
- Test path features over the knowledge graph to compute similarity between drugs

Thank you!