Chemical vs Network

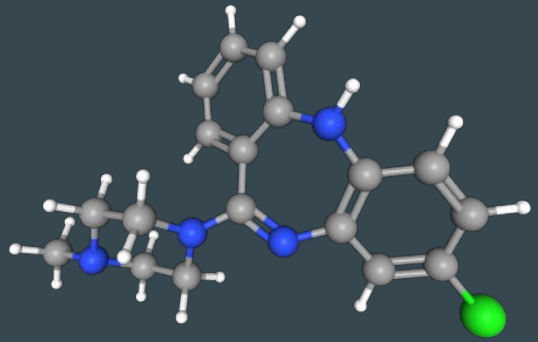
Information in Drug-Drug Interaction Prediction

George Kefalas, National Technical University of Athens
Dimitrios Vogiatzis, NCSR “Demokritos”

Main Approaches to DDIs Prediction

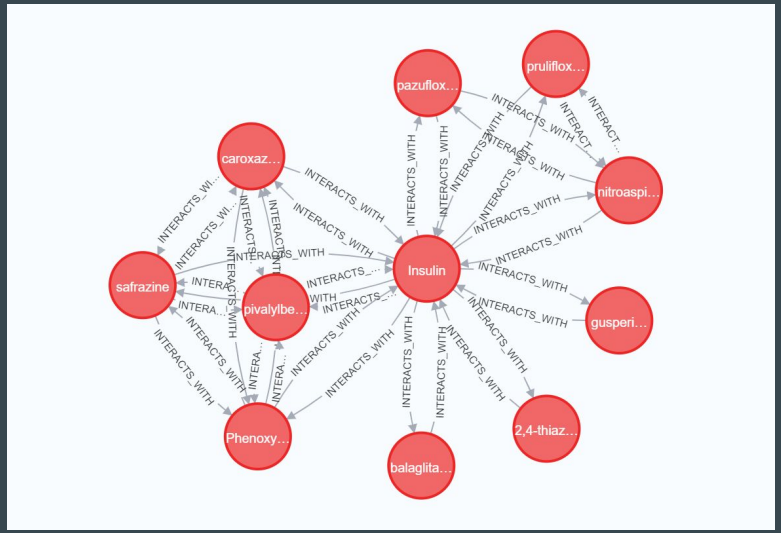
Chemical Information

- Molecular Structure
- Physical Properties
  - Physical state at room temperature
  - Water solubility
  - Melting and boiling temperatures
  - Electrical conductivity



Network Information

- Data Graph from known DDIs:
  - Drugs are denoted by nodes
  - Interactions are denoted by edges



Objective: Comparison between Chemical & Network Information

Comparison Base: Graph topological features ~ node degrees, k-core shells, betweenness centrality

Steps:

1. Harvest a diverse and reliable repository of DDIs: DrugBank
2. Produce vector embeddings from:
    - Chemical Information of each drug: mol2vec algorithm
    - Network Information of each drug: node2vec algorithm
3. Design & Train Classifiers: mol2vec, node2vec, hybrid
4. Study Classifiers' predictions over graph features:degree, core difference, centrality measures
5. Compare classifier behavior

DrugBank Repository (v.5.1.9)

- 14,624 Drugs
- 1,389,184 DDIs
- Unique Identifier for each drug
- Human curated → verified interactions

Drug Name	# Interactions	DrugBank Categories	DrugBank ID
Quinidine	2402	approved, investigational	DB00908
Clozapine	2379	approved	DB00363
Chlorpromazine	2369	approved, investigational vet_approved	DB00477

Data Harvest

- 3,753 Drugs
- 1,207,953 DDIs
- *full graph* for our experiments

Avg. Degree	643.73
Clustering Co.	0.621
Conn. Components	1
Diameter	5

Embedding Mechanisms

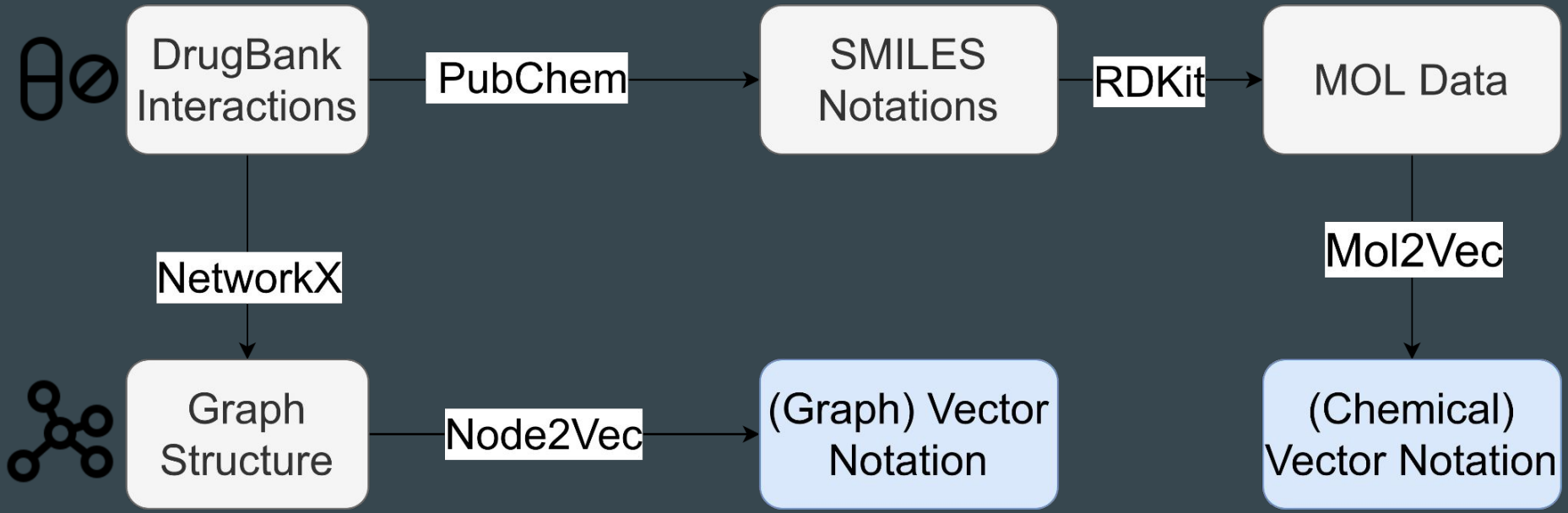
mol2vec

- Drugs format: SMILES notation
  C18H19ClN4 →
  CN1CCN(CC1)C2=NC3=C(C=CC(=C3)Cl)NC4=CC=CC=C42
- Drugs → low dimensional vector embeddings
- Unsupervised learning on all available chemical matter: **pre-trained** model for vectors of 300 dimensions.
- Similar vector embeddings denote drugs with similar molecular substructures.

node2vec

- Nodes → low dimensional vector embeddings
- Random walks on graph defined node neighbourhoods
- Embeddings follow likelihood maximization to preserve neighbourhood properties
- node2vec model is **trained on DrugBank** data before used to produce embeddings

Data Processing Pipeline



Experiment Design

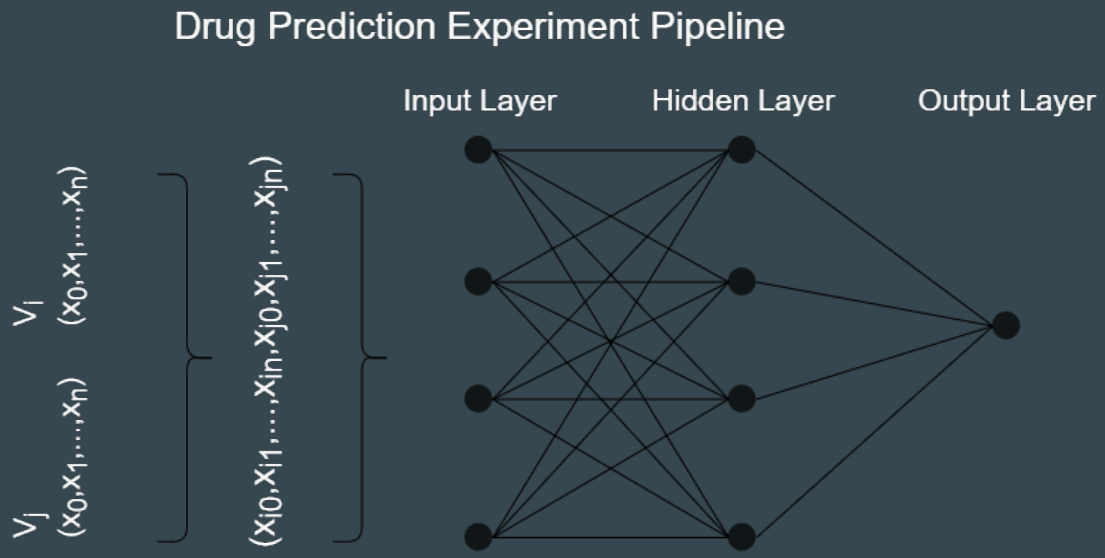
Classifiers:

- mol2vec
- node2vec
- hybrid

Input: concatenated drug (node) embeddings denote a training sample of a possible interaction.

Class 0: Negative Interaction
Class 1: Positive Interaction

Order of concatenation → two samples from each interaction



Issue no.1: node2vec embeddings

Training node2vec on the *full graph*, to embed the neighbourhood of each node, results in direct interactions being included in the vector representations as a key characteristic → no test set for node2vec classifier.

What we really need: adequate information to train the node2vec algorithm without (most) knowledge of direct DDIs.

Solution: Edge sampling on the *full graph* to produce the *sample graph*.

Sampling Target: keep all the nodes of the full graph and only 1% of its edges.

Sampled Graph

Graph	Nodes	Edges	Avg. Degree	Clustering Co.	Conn. Components	Diameter
Full	3,753	1,207,953	643.73	0.621	1	5
Sampled	3,753	12,080	6.44	0.005	799	∞

Table 1: Basic Properties for Full Graph and Sampled Graph

Graph	Degree C.	Eigenvector C.	Closeness C.	Betweenness C.
Full	0.1715	0.012	0.5030	0.00027
Sampled	0.0017	0.011	0.1556	0.00049

Table 2: Average Centrality Measures for Full Graph and Sampled Graph

Centrality Measures

- Closeness Centrality: average farness to all other nodes.

$$C(x) = \frac{N}{\sum_y d(y,x)}.$$

- Eigenvector Centrality: high-scoring nodes contribute greater score than low-scoring nodes.

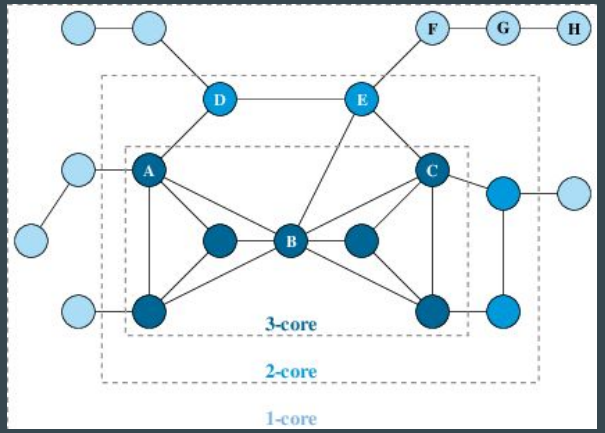
$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in V} a_{v,t} x_t$$

- Betweenness centrality: detecting the amount of influence a node has over the flow of information in a graph.

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

More Graph Measures

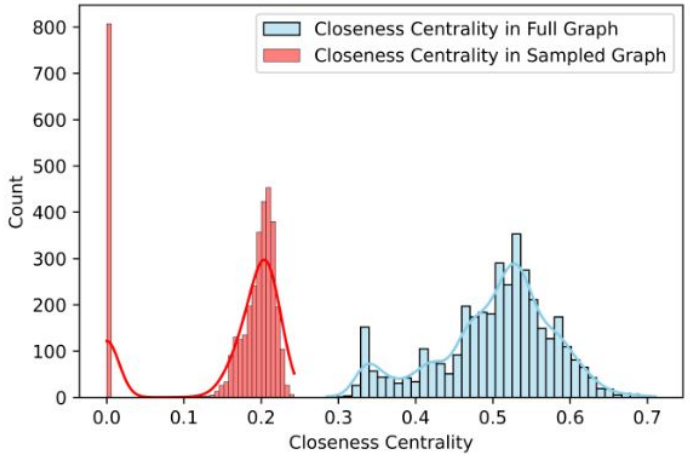
- K-Cores: A k-core of a graph G is a maximal connected subgraph of G in which all vertices have degree at least k.
- Core Value of nodes: A vertex v has coreness c if it belongs to a c-core but not to any (c+1)-core.



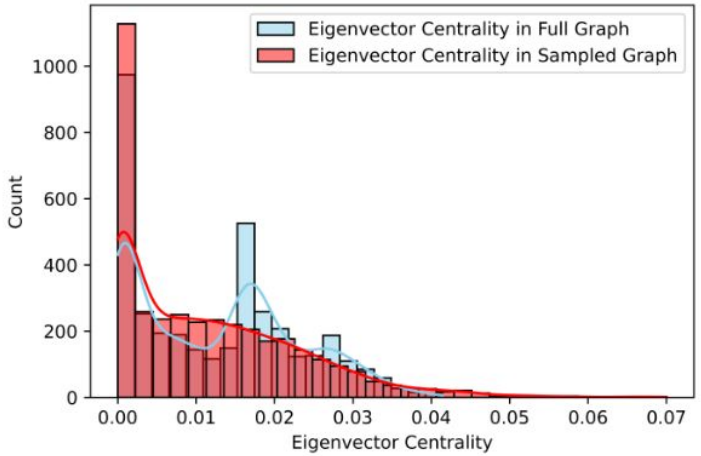
- **Core Difference** of a sample interaction:

$$\left| c_{v_1} - c_{v_2} \right|$$

But why 1%?



(a) Histogram and KDE of Closeness Centrality in *Full graph* and *Sampled graph*



(b) Histogram and KDE of Eigenvector Centrality in *Full Graph* and *Sampled Graph*

Fig. 3: Closeness Centrality & Eigenvector Centrality

Issue no.2: Negative Interactions

There are no negative interactions in DrugBank.

There is no datasource of negative interactions publicly available.

We create our own negative interactions through random sampling pairs of nodes. Samples that already constitute positive interactions are discarded and, similarly to positive interactions, samples produce two embeddings depending on the order of the concatenation.

Dataset

- Sample Graph is used to produce the node2vec embeddings

- Full Graph → Train/Validation/Test split: 65%-5%-35%

	Class	Train	Val	Test
mol2vec	0	1.502.158	266.288	870.700
	1	1.376.692	241.746	797.468
	Total	2.878.850	508.034	1.668.168
	Class	Train	Val	Test
node2vec	0	1.416.939	250.871	821.458
	1	1.036.784	182.140	1.196.982
	Total	2.453.723	433.011	2.018.440
	Class	Train	Val	Test
hybrid	0	1.417.479	250.615	821.598
	1	1.444.595	254.458	716.853
	Total	2.862.074	505.073	1.538.451

Table A.1. *Dataset split for each classifier*

Experimental Results (I)

Class	Classifier	Precision	Recall	F1 Score
Negative Samples (Class 0)	mol2vec	0.91	0.89	0.90
	node2vec	0.90	0.94	0.92
	Hybrid	**0.96**	**0.94**	**0.95**
Positive Samples (Class 1)	mol2vec	0.88	0.91	0.89
	node2vec	**0.96**	0.93	0.94
	Hybrid	0.93	**0.95**	**0.94**

(a) Classification Report

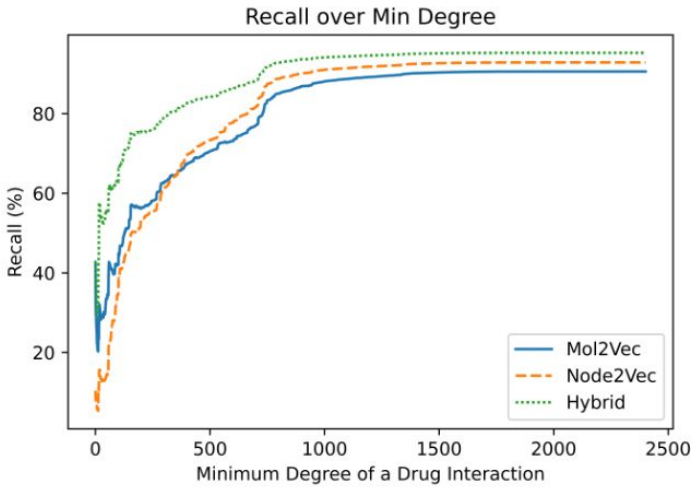
Category	mol2vec	node2vec	Hybrid
Experimental	87	94	96
Approved	85	93	94
Investigational	87	94	95
Vet Approved	88	94	97
Withdrawn	87	96	97
Illicit	87	95	96
Nutraceutical	91	90	97

(b) Accuracy percentage comparison for each drug category

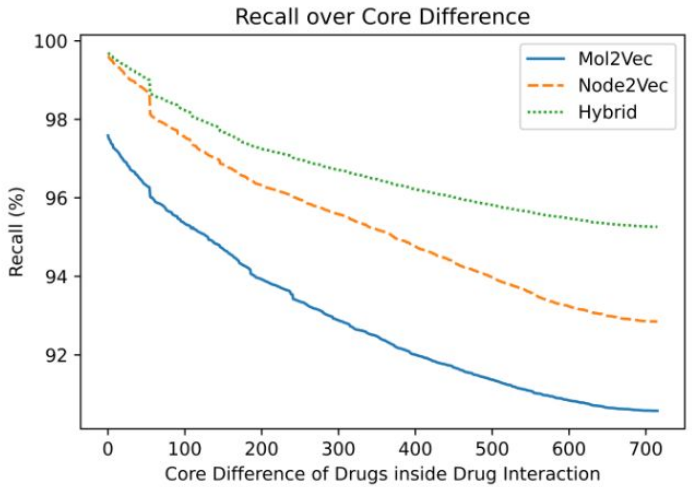
Fig. 5: Classification Report & Accuracy Comparison per Drug Category

- All models show notable performance
- Hybrid classifier outperforms mol2vec & node2vec → knowledge gain enough to counter larger vectors
- Exception: node2vec shows better precision on positive samples

Experimental Results (II)



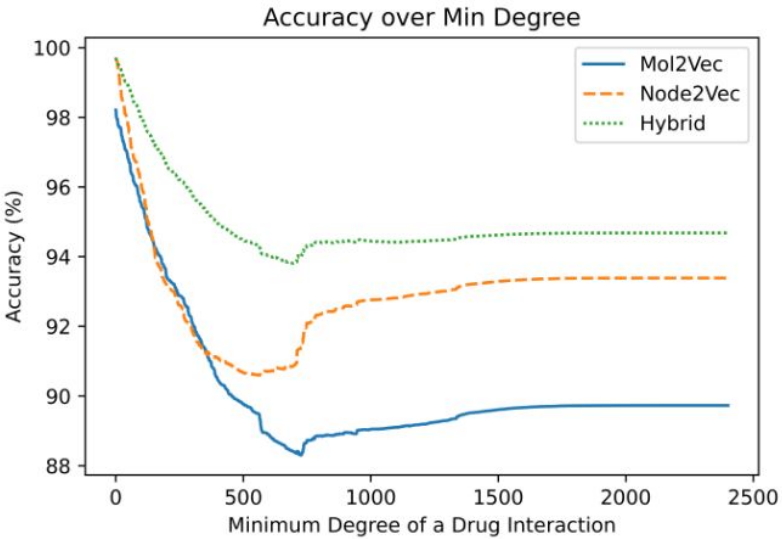
(a) Recall over Min Degree



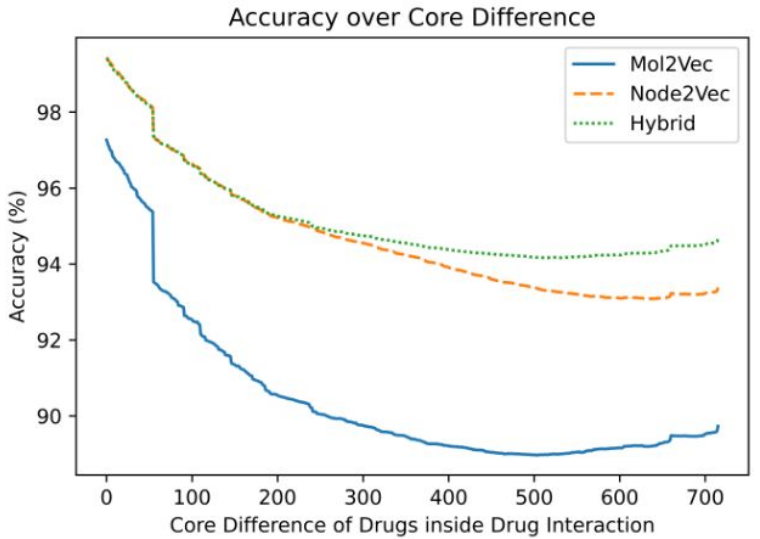
(b) Recall over Core Difference

Fig. 6: Recall plots for mol2vec, node2vec and Hybrid Classifiers

Experimental Results (III)



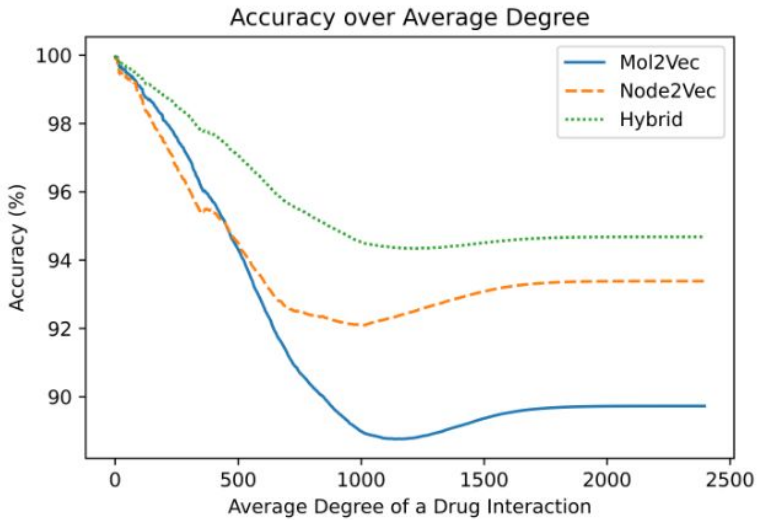
(a) Accuracy over Min Degree



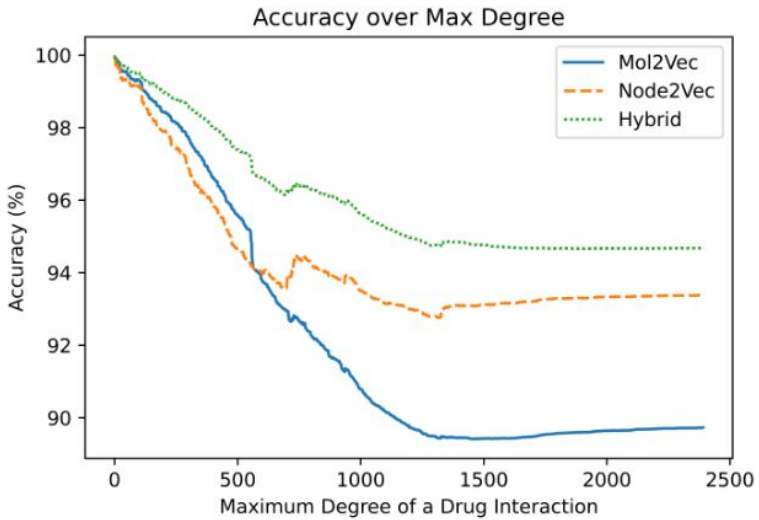
(b) Accuracy over Core Difference

Fig. 7: Accuracy plots for mol2vec, node2vec and Hybrid Classifiers

Experimental Results (IV)



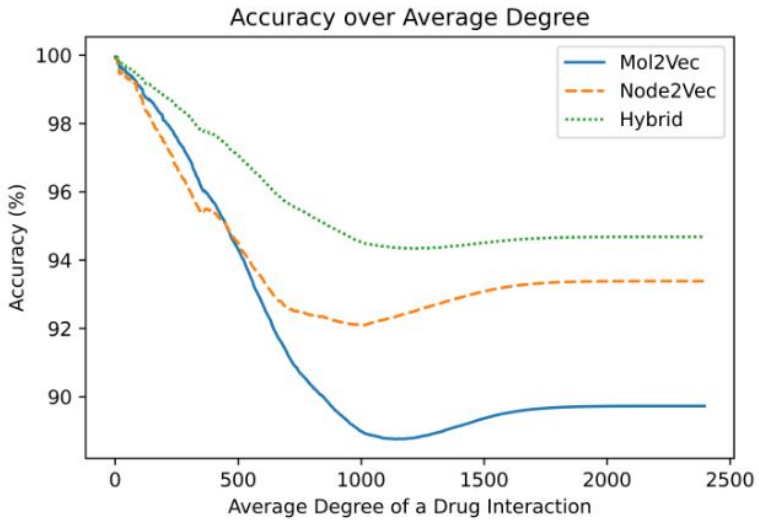
(a) *Accuracy over Avg. Degree*



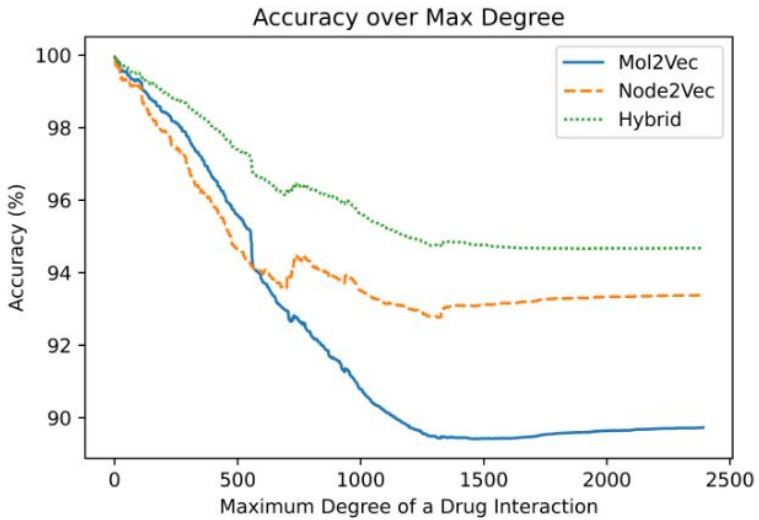
(b) *Accuracy over Max Degree*

Figure 4.6. *Accuracy plots for mol2vec, node2vec and hybrid Classifiers for avg. and max degree of sample interactions*

Experimental Results (IV)



(a) *Accuracy over Avg. Degree*



(b) *Accuracy over Max Degree*

Figure 4.6. *Accuracy plots for mol2vec, node2vec and hybrid Classifiers for avg. and max degree of sample interactions*

Experimental Results (V)

	Metric	mol2vec	node2vec	Hybrid
False Prediction	Average Min Degree	518.06	470.07	517.01
	Average Max Degree	1137.02	1143.09	1163.44
	Average Mean Degree	827.54	806.58	840.22
	Average Core Difference	237.17	266.10	242.34
	Average Betweenness Centrality	4.57e-06	6.99e-06	6.76e-06
Correct Prediction	Average Min Degree	519.66	593.45	512.09
	Average Max Degree	1083.25	1143.96	1078.67
	Average Mean Degree	801.45	868.713	795.38
	Average Core Difference	228.50	196.01	231.80
	Average Betweenness Centrality	1.39e-06	1.29e-06	1.44e-06

(a) Comparison of average graph properties of interactions grouped by each model's prediction

Classifier	Correct	False
Average Degree	0.018	0.047
Min Degree	0.019	**0.199**
Max Degree	0.010	**0.110**
Core Difference	0.013	**0.103**
Betweenness C.	**0.109**	0.071

(b) KL Divergence between mol2vec and node2vec sample distributions (values over 0.1 in bold)

Fig. 8: Result Metrics by Classifier Choice (Correct or False)

Conclusions (I)

Before we discuss the comparison of Chemical vs Network Information:

- DrugBank is a useful data set for DDIs prediction experiments
- node2vec application: extreme efficiency in discovering neighbourhoods from only a fraction of a graph & great tool for predicting drug interactions
- mol2vec gamble: scarce examples of usage for DDIs prediction. Although its main application is drug discovery, we found it to also be efficient in interaction prediction.

Conclusions (II)

- hybrid > node2vec > mol2vec – Utilizing both chemical and network embeddings leads to better prediction → not overlapping information
- mol2vec shows better recall than node2vec for lower min. degrees, threshold = 350
- lower core difference in samples is correlated with better efficiency in all models
- node2vec reaches performance equal to hybrid model for lower core difference

Future Work

- Multi-Label Prediction

- Efficient Construction of Negative Samples
  - e.g. Positive Corruption, Nearest Neighbours

- Augmented Graph (e.g. protein nodes)

Thank You