

# Gyógyszer kölcsonhatások azonosítása gráf autoenkóderrel

BILIBÓK BENCE, PÉTER ISTVÁN, RANCZ MÁTÉ

# A téma motivációja

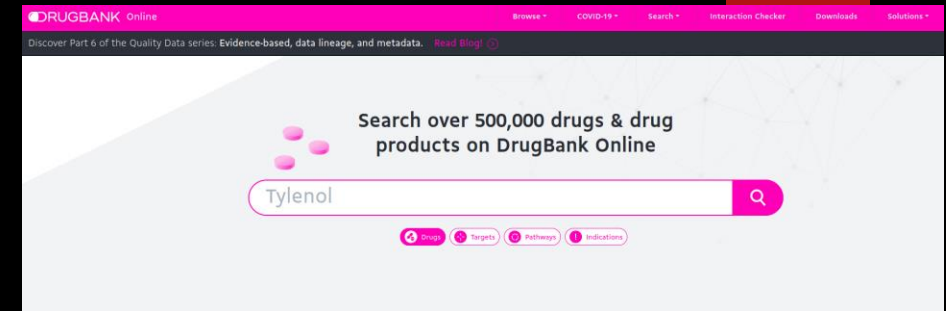
- ▶ A gyógyszeriparban egy hatóanyag biztonságossága szempontjából fontos, hogy milyen más hatóanyagokkal lép kölcsönhatásba
- ▶ Ezek az *adverz* reakciók, legyen erősítő vagy gyengítő
- ▶ Többnyire kísérleti módszerek
- ▶ DL alapú predikció

# Probléma megfogalmazása

- ▶ Élpredikció gráfon
- ▶ A csúcsok a hatóanyagok
- ▶ Az élek adverzreakciók, ha vannak

# Adatok

- ▶ BioSNAP, Drug-drug
- ▶ 1514 hatóanyag, DrugBank azonosítóval
- ▶ kb. 48k él
- ▶ DrugBank



By Jure Leskovec

STANFORD UNIVERSITY

## Drug-drug interaction network

**Dataset information**

This is a network of interactions between drugs, which are approved by the U.S. Food and Drug Administration. Nodes represent drugs and edges represent drug interactions. Drug-drug interactions occur when the pharmacologic effect of a one drug is altered by the action of another drug, leading to unpredictable clinical effects such as adverse drug reactions. When several drugs are administered together, there might be a greater possibility of adverse drug reactions as one drug can increase or decrease the effect of another drug. It is thus crucial to identify drug-drug interactions, especially during drug development, when one needs to identify interactions between a new candidate drug and drugs that are already on the market.

**Dataset statistics**

Nodes	1514
Edges	48514
Nodes in largest SCC	1510
Fraction of nodes in largest SCC	1.000000
Edges in largest SCC	48512
Fraction of edges in largest SCC	0.999959
Average clustering coefficient	0.303968
Number of triangles	5457046
Fraction of closed triangles	0.104354
Diameter (longest shortest path)	7
90-percentile effective diameter	2.977804

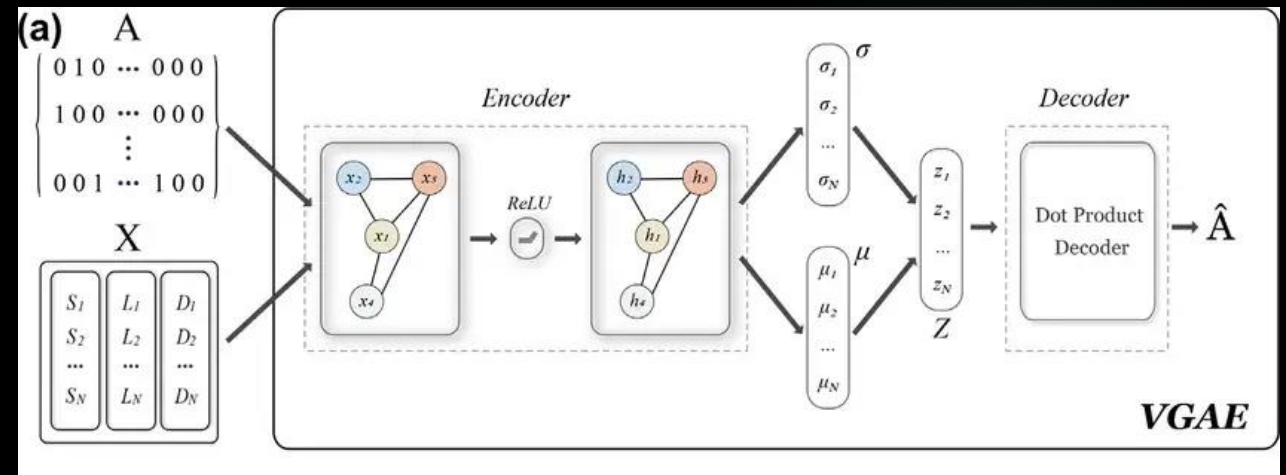
**Open positions**

Open research positions in SNAP group are available [here](#).

SNAP for C++  
SNAP for Python  
SNAP Datasets  
BIO-SNAP Datasets  
What's new  
People  
Papers  
Projects  
Using SNAP  
Links  
About  
Contact us

# Modell: VGAE

- ▶ Variációs Gráf Autoenkóder
- ▶ Gráf konvolúciós réteg
- ▶ Látens reprezentáció
- ▶ Mintavételezés egy normál eloszlásból
- ▶ Két csúcs látens címkéi közti viszony határozza meg, hogy össze lehetnek-e kötve
- ▶ Skaláris szorzat



# Modell: VGAE

**Inference model** We take a simple inference model parameterized by a two-layer GCN:

$$q(\mathbf{Z} | \mathbf{X}, \mathbf{A}) = \prod_{i=1}^N q(\mathbf{z}_i | \mathbf{X}, \mathbf{A}), \quad \text{with} \quad q(\mathbf{z}_i | \mathbf{X}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i | \boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)). \quad (1)$$

Here,  $\boldsymbol{\mu} = \text{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$  is the matrix of mean vectors  $\boldsymbol{\mu}_i$ ; similarly  $\log \boldsymbol{\sigma} = \text{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$ . The two-layer GCN is defined as  $\text{GCN}(\mathbf{X}, \mathbf{A}) = \tilde{\mathbf{A}} \text{ReLU}(\tilde{\mathbf{A}} \mathbf{X} \mathbf{W}_0) \mathbf{W}_1$ , with weight matrices  $\mathbf{W}_i$ .  $\text{GCN}_{\boldsymbol{\mu}}(\mathbf{X}, \mathbf{A})$  and  $\text{GCN}_{\boldsymbol{\sigma}}(\mathbf{X}, \mathbf{A})$  share first-layer parameters  $\mathbf{W}_0$ .  $\text{ReLU}(\cdot) = \max(0, \cdot)$  and  $\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$  is the symmetrically normalized adjacency matrix.

**Generative model** Our generative model is given by an inner product between latent variables:

$$p(\mathbf{A} | \mathbf{Z}) = \prod_{i=1}^N \prod_{j=1}^N p(A_{ij} | \mathbf{z}_i, \mathbf{z}_j), \quad \text{with} \quad p(A_{ij} = 1 | \mathbf{z}_i, \mathbf{z}_j) = \sigma(\mathbf{z}_i^T \mathbf{z}_j), \quad (2)$$

where  $A_{ij}$  are the elements of  $\mathbf{A}$  and  $\sigma(\cdot)$  is the logistic sigmoid function.

**Learning** We optimize the variational lower bound  $\mathcal{L}$  w.r.t. the variational parameters  $\mathbf{W}_i$ :

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z} | \mathbf{X}, \mathbf{A})} [\log p(\mathbf{A} | \mathbf{Z})] - \text{KL}[q(\mathbf{Z} | \mathbf{X}, \mathbf{A}) || p(\mathbf{Z})], \quad (3)$$

# Jellemzők kinyerése

- ▶ Adatbázis vizsgálata
- ▶ Sok jellemző szöveggel van leírva -> nem megfelelő nekünk
- ▶ Számszerű jellemzők kiválasztása (tömeg, olvadási pont, stb.)
- ▶ Molekuláris összetétel és osztályozások -> one-hot kódolással

# Tanítás

- ▶ <https://github.com/dmlc/dgl/tree/master/examples/pytorch/vgae>
- ▶ Binary cross entropy (élpredikció) + KL divergence (látens eloszlás)
- ▶ Az élek 5 ill. 10%-a teszt és validációs él (maszkolás a tanítás során)
- ▶ <https://github.com/peter-i-istvan/nhf-drugs>
- ▶ 0.05 learning rate, Adam
- ▶ 10 hosszú bemeneti vektor (feature), mint címke
- ▶ 32 dimenziós látens tér



# Eredmények, kiértékelés

