# Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules

胡晨琪 马佳良 刘群

Center for Earth System Science

Paper Authors: Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi
Published in *Journal of Chemical Information and Modeling*, 2013.

June 8, 2015

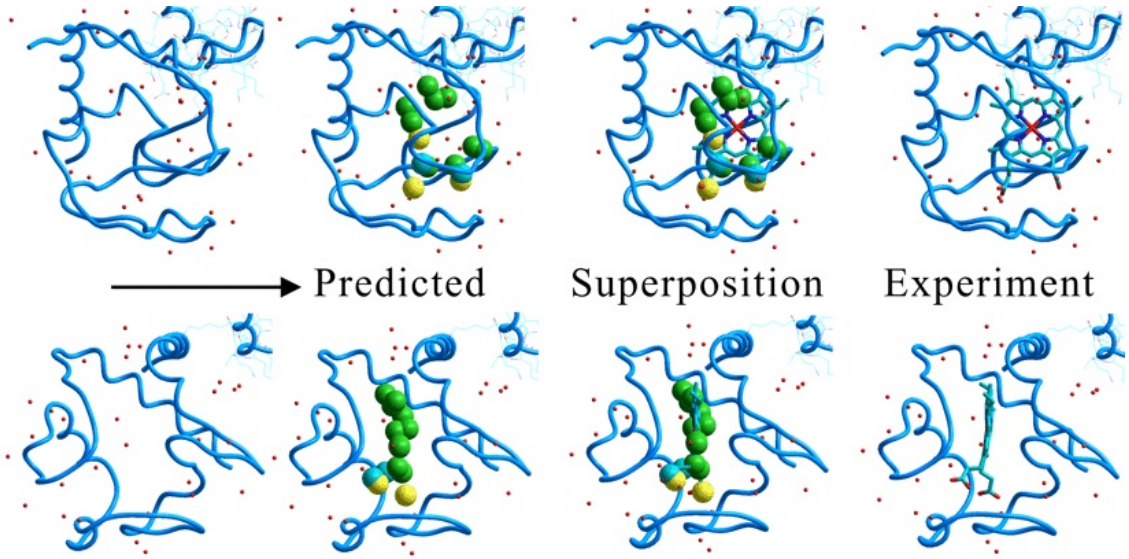
# Outline

# Background

- Aqueous solubility prediction is important to drug discovery.
- Original method: QSAR (Quantitative Structure-Activity Relationship) methods

$$Activity = F(structure) = M(E(structure))$$

$E$ : Encoding function  $M$ : Mapping function




(http://www.molfunction.com/software5.htm)

# Background - Deep Learning

1. How about autoencoder-based and convolutional architectures?
   - Molecular properties should be represented by vectors of fixed length.
   - Rely heavily on a good encoding function.
2. Molecules are naturally represented by small graphs of *variable size*.

# Background - Deep Learning

1. How about autoencoder-based and convolutional architectures?
   - Molecular properties should be represented by vectors of fixed length.
   - Rely heavily on a good encoding function.
2. Molecules are naturally represented by small graphs of *variable size*.

So how should we do?

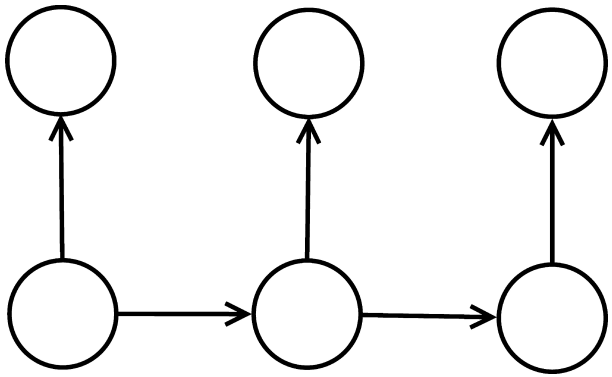
# Background - Deep Learning

1. How about autoencoder-based and convolutional architectures?
   - Molecular properties should be represented by vectors of fixed length.
   - Rely heavily on a good encoding function.
2. Molecules are naturally represented by small graphs of *variable size*.

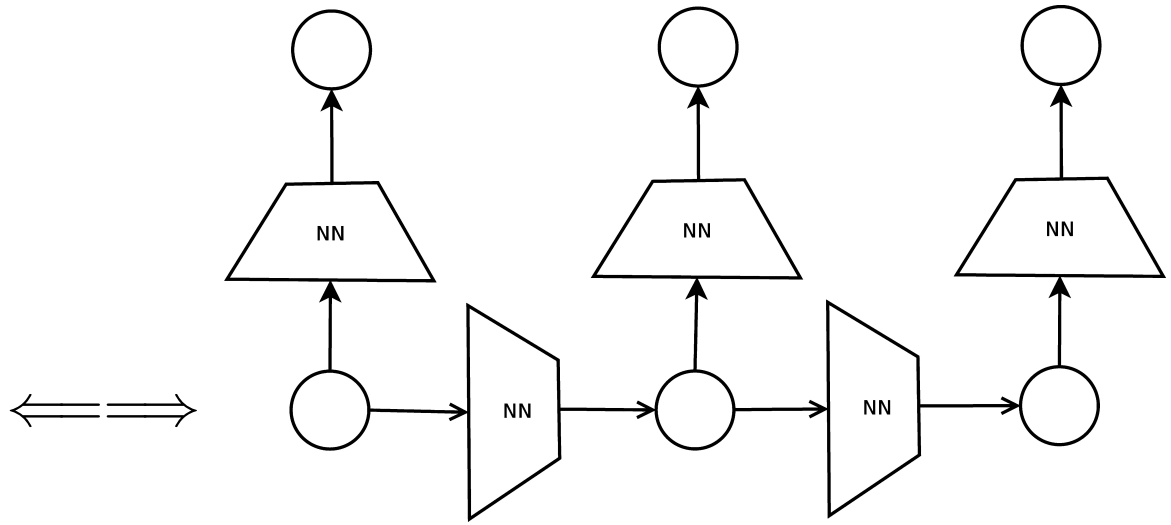
So how should we do?

- Directed acyclic graph recursive neural networks

# Directed Acyclic Graph Recursive Neural Networks



Directed acyclic graph(DAG)

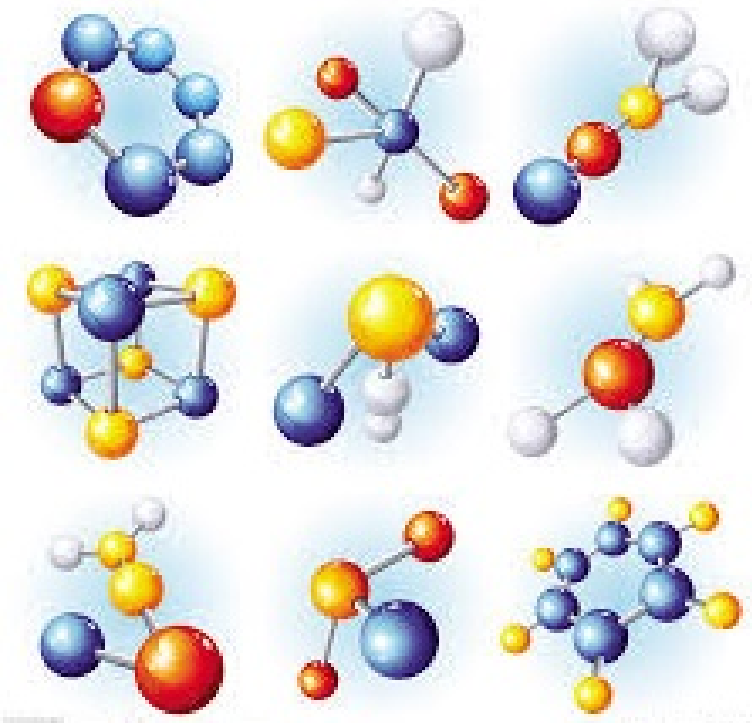


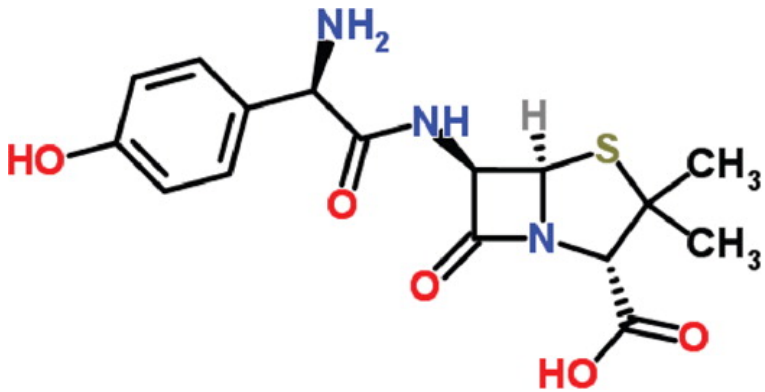

Directed acyclic graphs recursive neural netw
(DAG-RNN)

The DAG-RNN approach associates vector variables with the nodes of the DAG and places a neural network (or any other kind of parametrized function) on the edges of the DAG to parametrize the relationship between the corresponding vector variables.

# Question

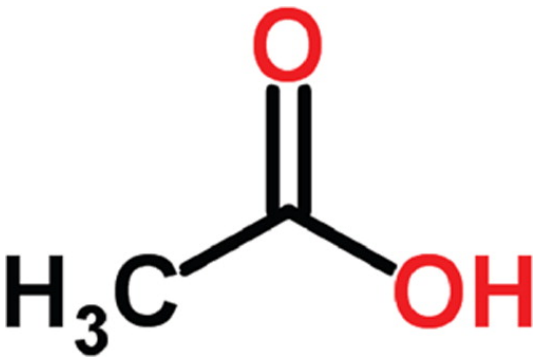
But the molecules are undirected graphs(UG) and possibly cyclic, how could we use the DAG-RNN architecture?

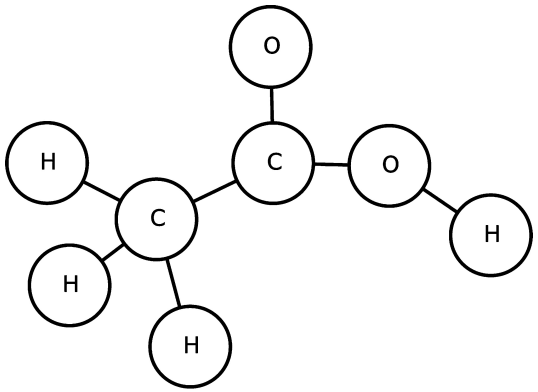


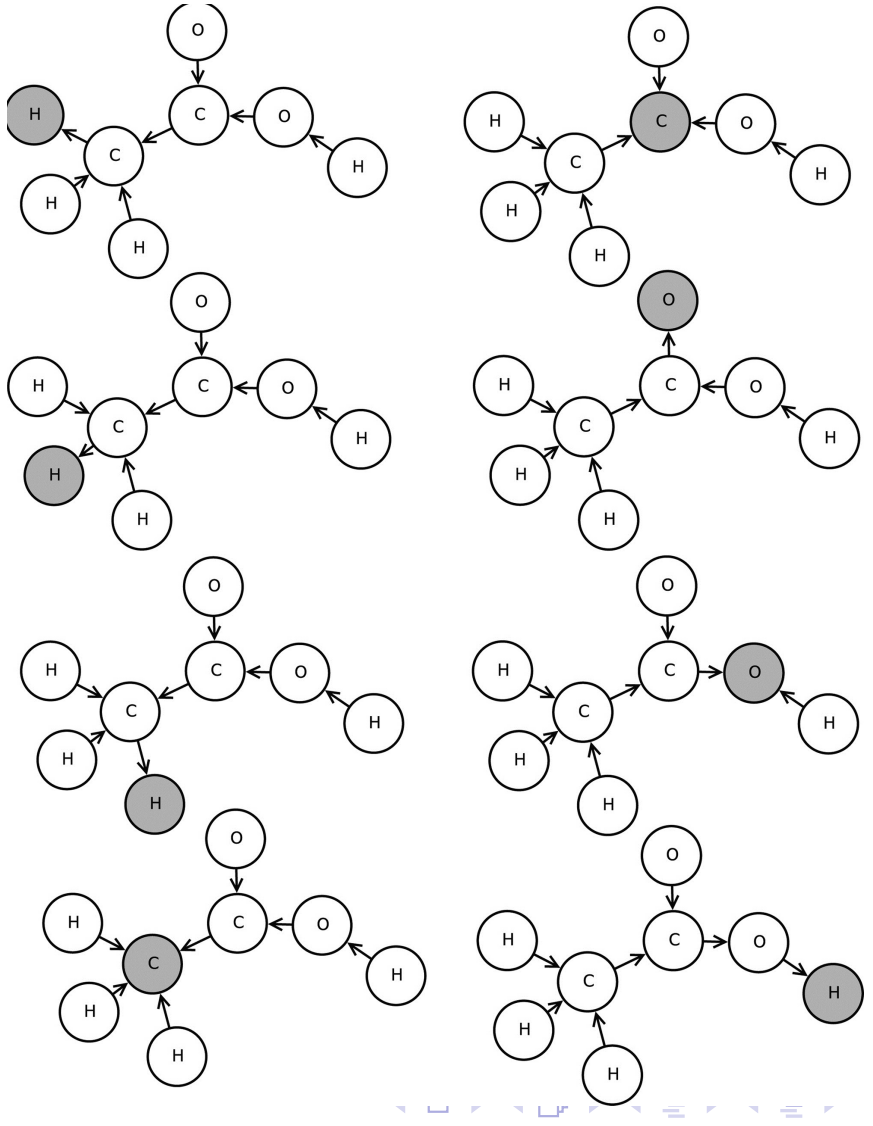



Amoxicillin

# Undirected Graph Recursive Neural Networks (UGRNN)

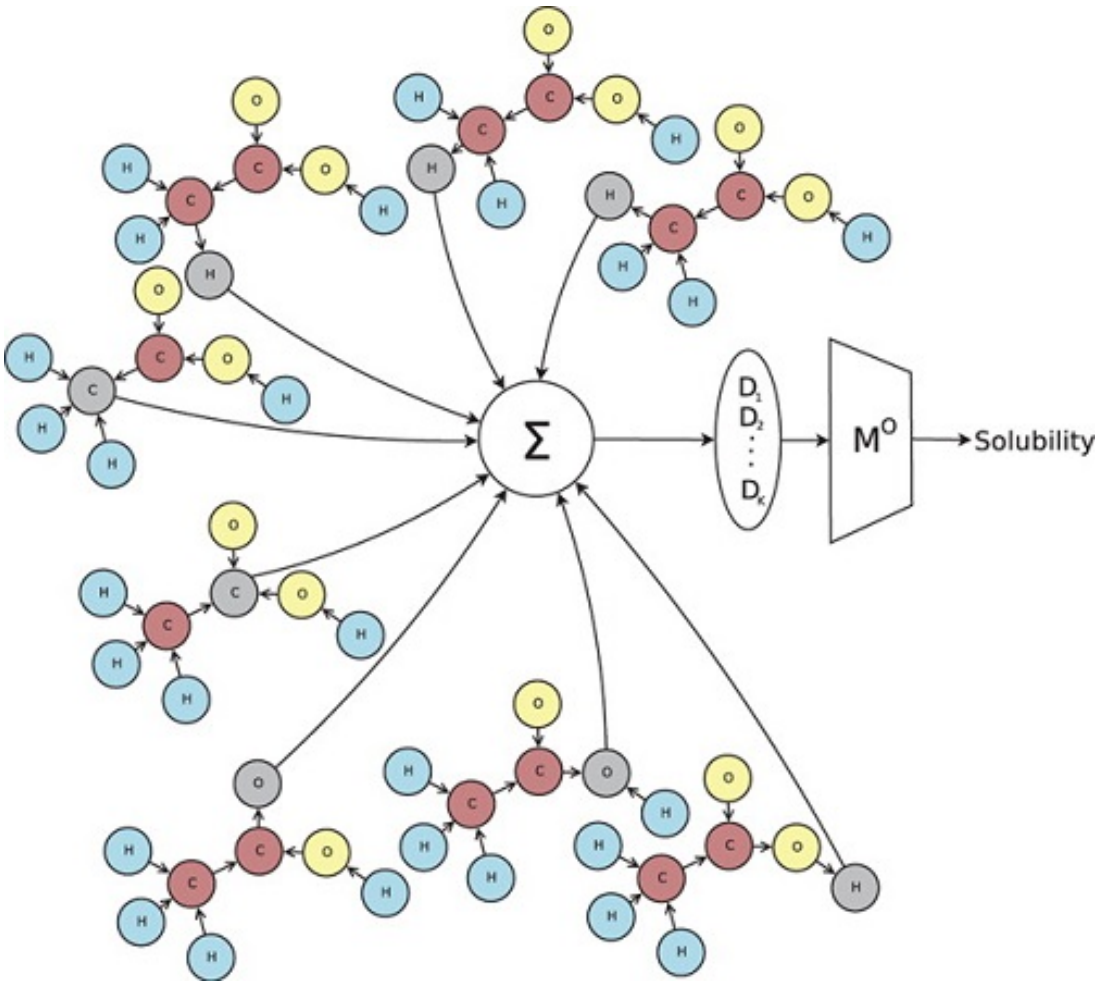


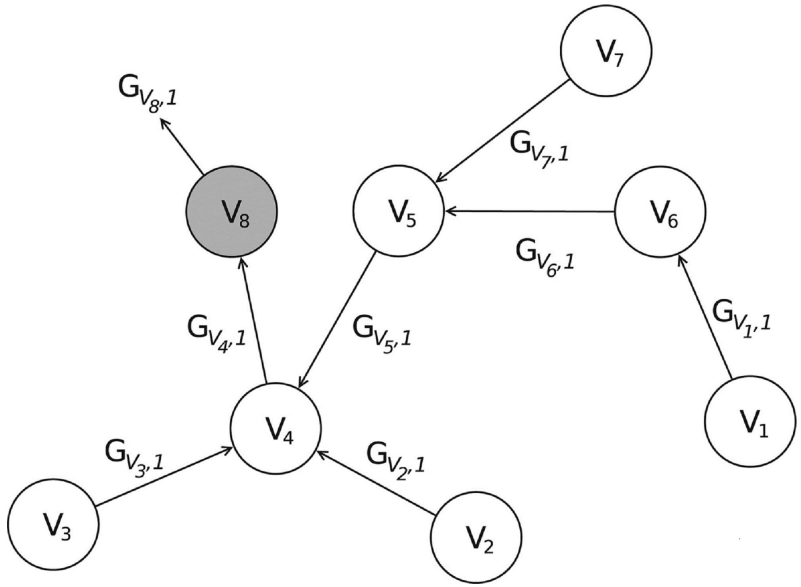




etic acid and its undirected graph

# Undirected Graph Recursive Neural Networks (UGRNN)




Sum of eight $G$ vectors to produce the vector $G_{\text{structure}} = (D_1, \ldots, D_K)$ corresponding to $K$ descriptors learned from the data. The output function $M^O$ produces the final prediction.




*Equations* :

$$G_{v,k} = M^G(i_v, G_{pa^1_{[v,k]}}, \cdots, G_{pa^n_{[v,k}}$$

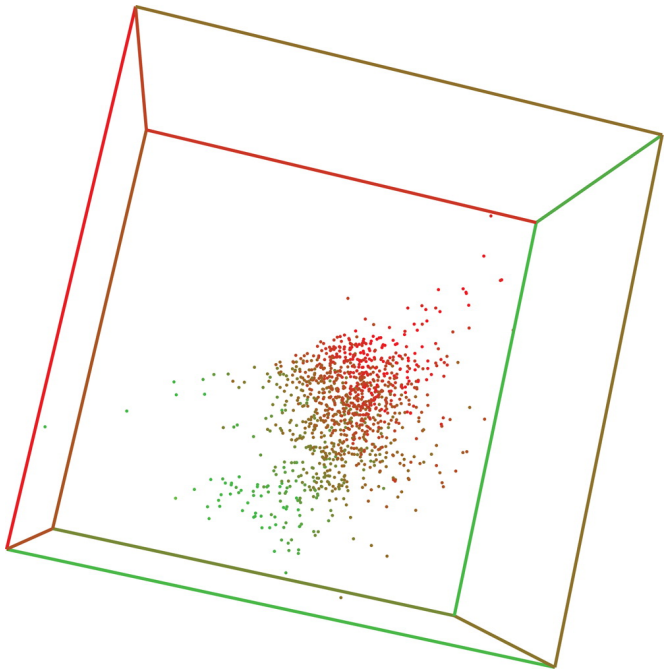
$$G_{\text{structure}} = \sum_{k=1}^{N} G_{r_k,k} = (D_1, \cdots$$

$$p = M^O(G_{\text{structure}})$$

# Data and Results

Prediction Performances and Standard Deviations Using 10-Fold Cross Validation on the Small Delaney Data Set

| models | $R^2$ | std $R^2$ | RMSE | std RMSE | AAE | std AAE |
|---|---|---|---|---|---|---|
| UG-RNN | 0.92 | 0.02 | 0.58 | 0.07 | 0.43 | 0.04 |
| UG-RNN-CR | 0.86 | 0.03 | 0.79 | 0.09 | 0.57 | 0.06 |
| UG-RNN + log P | 0.91 | 0.02 | 0.61 | 0.07 | 0.46 | 0.05 |
| UG-RNN-CR + log P | 0.91 | 0.02 | 0.63 | 0.05 | 0.47 | 0.03 |
| GSE(23) | – | – | – | – | 0.47 | – |
| 2D kernel(param d = 2) | 0.91 | – | 0.61 | – | 0.44 | – |



Scatter plot of learned feature vectors for molecules in the small Delaney data set.

# Discussion and Conclusions

The performance of the deep learning methods matches or exceeds the performance of other state-of-the-art methods according to several evaluation metrics and expose the fundamental limitations arising from training sets that are too small or too noisy.

# Q&A

Q & A

# The End

Thank You!