N-Gram Graph, A Novel Molecule Representation

Shengchao Liu[1], Thevaa Chandereng[2], Yingyu Liang[1]

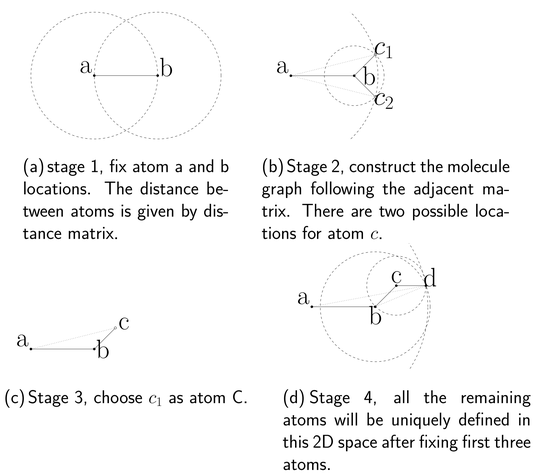
[1]Department of Computer Science, University of Wisconsin-Madison; [2] Department of Statistics, University of Wisconsin-Madison

Objectives

Virtual high-throughput screening provides a strategy for prioritizing compounds for physical screens. Machine learning methods offer an ancillary benefit to make molecule predictions, yet the choice of representation has been challenging when selecting algorithms. We emphasize the effects of different levels of molecule representation. Then, we introduce N-Gram graph, a novel representation for a molecular graph. We demonstrate that N-Gram graph is able to attain most accurate prediction with several non-deep machine learning methods on multiple tasks.

Motivation

- Message passing based on adjacent matrix can help identify a molecule skeleton.
- Distance matrix maintains the information of a molecule shape.
- Combining both can keep all the key information in a molecule.



(a) stage 1, fix atom a and b locations. The distance between atoms is given by distance matrix.

(b) Stage 2, construct the molecule graph following the adjacent matrix. There are two possible locations for atom c.

(c) Stage 3, choose c_1 as atom C.

(d) Stage 4, all the remaining atoms will be uniquely defined in this 2D space after fixing first three atoms.

Figure 1: Illustrations on how adjacent matrix and distance matrix can be combined to recover a graph structure.

Graph Representation

Each molecule can be represented as a graph with at most m atoms. Each atom can be represented as a vector of $d-$dimension.

- Adjacent Matrix $\mathcal{A} \in \{0,1\}^{m\times m}$

$$\mathcal{A}_{i,j} = \begin{cases} 1, & \text{atom}_i \text{ and atom}_j \text{ are bonded} \\ 0, & \text{otherwise} \end{cases}$$

- Distance Matrix $\mathcal{D} \in \mathbb{R}^{m\times m}$

$$\mathcal{D}_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

- Node Attribute Matrix $\mathcal{N} \in \{0,1\}^{d\times m}$. For each atom, the features are symbol, degree, # Hytrogeon, charges, is aromatic, is acceptor, is donor.

$$\mathcal{N}_{\cdot,i} = [\underbrace{\text{C}, \text{Cl}, \text{I}, \text{F}, \ldots}_{\text{atom symbol}}, \underbrace{0, 1, 2, 3, 4, 5, 6}_{\text{atom degree}}, \ldots]$$

Methods: N-Gram Graph

Segmented Random Projection:

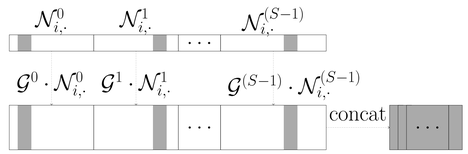


Figure 2: Segmented random projection on atom a_i. Each atom features can be split into S segments. Each group of feature with dimension d_s corresponds to a one-hot vector $N_{i,\cdot}^s \in \{0,1\}^{1\times d_s}$ (marked in grey). Multiply it by Gaussian random matrix $\mathcal{G}^s \in \mathbb{R}^{r\times d_s}$ as projection to a random space. For each randomized atom feature g_i, the only non-zero column in output matrix $\mathcal{G}^s \cdot \mathcal{N}_{i,\cdot}^s$ in each segment will be extracted and concatenated.

N-Gram Path and N-Gram Graph:

Let V be a path, and N-Gram path (V_n) is the production of all n nodes in that path. Let $\mathcal{V}_n \in \mathbb{R}^{r\times S}$, $p \in \{1, 2, \ldots, N\}$ represent the **N-Gram path set**. It is defined as the sum of all N-Gram paths with length n.

$$\mathcal{V}_n = \underbrace{\sum_{\forall V, \text{s.t. } |V|=n} \overbrace{\prod_{a_i \in V} \underbrace{f(a_i)}_{\text{segmented random projection}}}^{\text{N-Gram path}}}_{\text{n-graph path set}}$$

N-Gram graph for each molecule $\mathbb{G} = [\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_n] \in \mathbb{R}^{N\times r\times S}$ is the concatenation of N-Gram path sets with multiple length n.

Experiments

- Three regression tasks, Delaney, Malaria, and CEP.
- Six models are tested: RF, XGB, DNN, NEF [1], GCNN [2], Weave Net [3].

Table 1: RMSE on three regression tasks (test set). Top three results are **bolded** and the best performance is **underlined**. Baseline results (*) are from [1, 3].

Representation	Method	Delaney	Malaria	CEP
ECFP	RF	1.251	**1.011**	1.667
	XGB	1.120	**0.998**	1.442
	DNN (*)	1.40	1.13	2.00
Message-Passing Graph	NEF (*)	**0.52**	1.15	1.43
	GCNN	0.98	1.02	**1.17**
	Weave (*)	**0.46**	1.07	**1.10**
N-Gram Graph	RF	0.802	**1.011**	1.367
	XGB	0.771	**1.003**	**1.296**
	DNN	**0.665**	1.085	1.359

Conclusion and Discussion

- No requirement for End-to-End deep neural networks.
- Current graph-based methods haven't fully utilized the comprehensive capacity of deep neural network.
- More advanced NLP strategies can be applied for both modeling and analysis.

References

[1] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alan Aspuru-Guzik, and Ryan P Adams. Convolutional Networks on Graphs for Learning Molecular Fingerprints. pages 2224–2232, 2015.

[2] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4):283–293, 2017.

[3] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.