Learning Molecular Representation

Hy Truong Son, Nguyen Duc Hai
Advisor: Prof. Risi Kondor

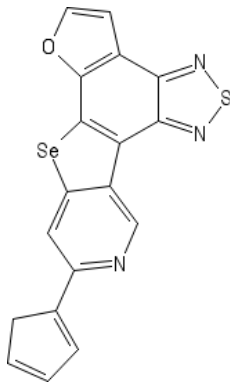
The University of Chicago

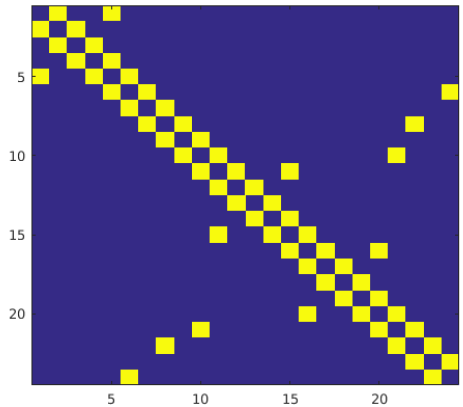
December 2017

THE UNIVERSITY OF CHICAGO

Harvard Clean Energy Project Dataset (HCEP)

SMILES: C1C=CC=C1c1cc2[Se]c3c4occc4c4nsnc4c3c2cn1
Power Conversion Efficiency (PCE, range 0 - 11): 5.16195





$C_{18}H_9N_3OSSe$ Adjacency matrix

THE UNIVERSITY OF CHICAGO

dK-Series Algorithms (2)

Reference:

Systematic Topology Analysis and Generation Using Degree Correlations
ACM SIGCOMM 2006

Algorithm:

1. Extract all subgraphs of size d in a molecular graph.
2. Classify each subgraph as 1 element of the set of non-isomorphic graphs of size d.
3. Based on the subgraph classification, we build the frequency vector (probability distribution) and use the frequency vector as the molecular graph representation.

Remark: Instead of using vertex degree, we use atomic types (for example, Carbon - C, Hydrogen - H, etc.).

THE UNIVERSITY OF CHICAGO

Results & Future Plan

Results of using Linear Regression on dK-features:

	Test MAE	Test RMSE
$d = 1$	1.492278	3.457323
$d = 2$	1.255755	2.545540
$d = 3$	1.130819	2.194515
$d = 4$	**1.025007**	**1.774689**

Next steps:

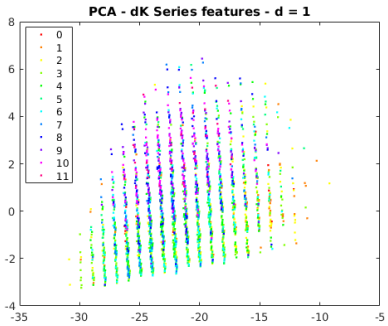
1. Apply more physical features rather than just atomic types
2. Apply state-of-the-art graph neural networks in learning molecular representations
3. Test on other molecular datasets

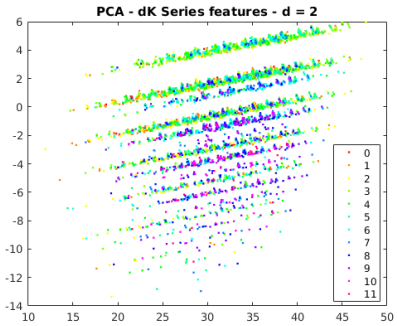
THE UNIVERSITY OF CHICAGO

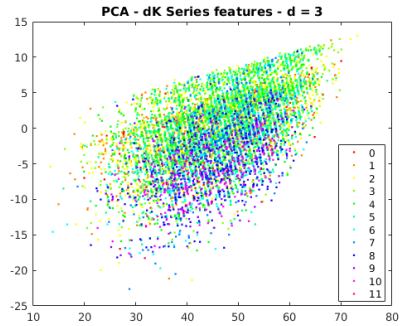
Thank you very much for your attention!

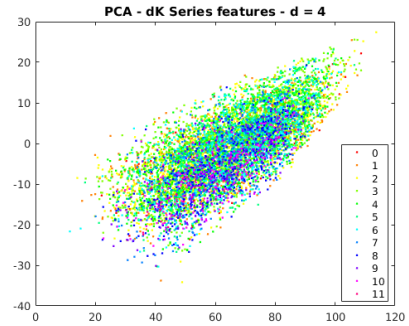
THE UNIVERSITY OF CHICAGO

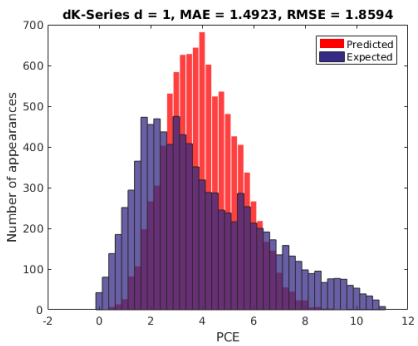
Q & A

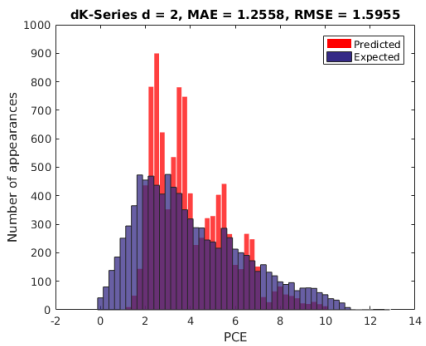


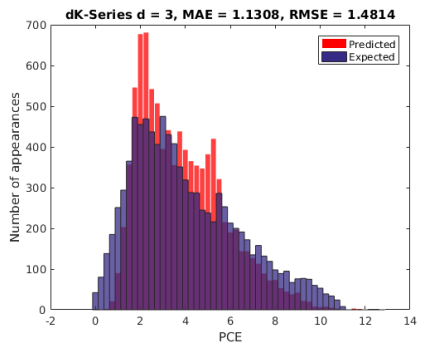


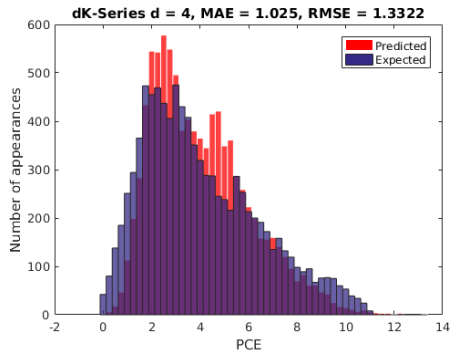












`https://github.com/HyTruongSon/dk-series`

For more detail, read our paper **Covariant Compositional Networks For Learning Graphs** (ICLR 2018 - Workshop) [Kondor et. al.]

THE UNIVERSITY OF CHICAGO