OrgoNet: Organic Chemistry Reaction Prediction

Ethan Chi, Bowen Jing, Emily Wen

ethanchi@cs.stanford.edu, bjing@cs.stanford.edu, ewen22@stanford.edu

Problem

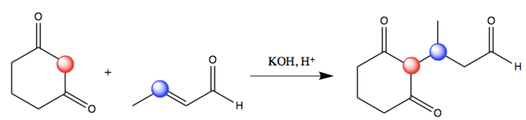
Given two molecules and a set of reagents, predict the product that will form.

Background

- Reaction prediction is a major problem in organic chemistry usually solved through expensive experimental methods.
- Current state of the art is *reaction templates*: non-generalizable subgraphs created by human input. [1]

Data

- 50,000 reactions from the US Patent Office database [split 80%-10%-10%]
- Preprocessed using chemical heuristics to identify *source* (electron donator) and *sink* (electron acceptor) molecules



Michael addition of cyclohexane-1,3-dione and (E)-but-2-enal to form 3-(2,6-dioxocyclohexyl)butanal.
Blue is sink; red is source.

References

[1] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). A comprehensive survey on graph neural networks. *arXiv preprint arXiv: 1901.00596*

[2] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of organic reaction outcomes using machine learning," *ACS Central Science*, vol. 3, no. 5, pp. 434–443, 2017.

Methods

1) Logistic Regression

Each atom is embedded using a hand-tuned naïve weighting algorithm that decays by 4x for every further distance from the atom. We then model the probability of a bond forming between a source atom a_i and a sink atom a_j under reagent k as the following:

$$P((a_i, a_j), R_k) = g(x_i^{\mathrm{T}} A_k x^j)$$

We want to learn A_k for each reagent k in the dataset. This model corresponds to a chemical intuition that each reagent corresponds to a *linear transformation* which maps source atom embeddings onto the direction of sink atom embeddings.

2) Gaussian Discriminant Analysis

Using the same embeddings as in logistic regression, we assume that bonding pairs $(x_i,\ x_j)$ are generated according a multivariate Gaussian with a Wishart prior. Since chemically, reactivity typically depends on the difference in electrical characteristics between two atoms, rather than their absolute values, we model the differences themselves:

$$x_i - x_j \sim \mathcal{N}(\mu_1, \Sigma_1)$$

3) Graphical Convolutional Neural Network (GNN)

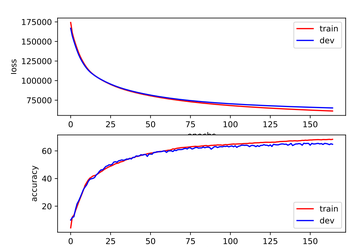
To embed atoms, we implement a graphical convolutional neural network as described in [2]. In layer t, we update the embedding of a node x_i as follows:

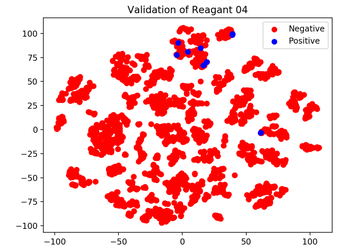
$$x_i^{[n]} := g\left(W_t^{[n]} x_i^{[n-1]} + \sum_{t \in T} \sum_{\{a_i, a_j\} \in V_t} W_t^{[n]} x_j^{[n-1]}\right)$$

We then apply the same linear transformation as in logistic regression for prediction.
We train with learning rate = 5e-06 and with a modified log loss: $LL = 5y \log \hat{y} + (1-y)\log(1-\hat{y})$

Results

Reagent	Log. Reg.	GDA	GNN
00	48.5%	16.4%	**65.3%**
01	42.7%	15.7%	**62.9%**
02	**60.0%**	22.1%	59.7%
03	59.7%	18.1%	**69.4%**
04	**83.2%**	30.4%	78.2%
Total	52.2%	17.9%	**65.2%**





Discussion

- Generally speaking, GNN performs better than log. reg., which performs better than the GDA.
- Log. reg. occasionally performs better for reactions for which the naïve embedding is well-suited (typically strongly acidic or basic reagents, e.g. HCl). However, the naïve manual embedding also makes performance heterogenous.
- GDA uniformly performed poorly, possibly due to a Gaussian being a poor model.

Future Work

- Transfer learning: using already-learned embeddings to speed up learning on future reagents
- More sophisticated GNN with bonds as nodes; also, more complex spectral methods for faster learning