WELCOME TO THE

Molecular Team Lecture Series

In this lecture series, MAI LAB Molecular Team will introduce various molecular design tasks

TODAY'S LECTURE

A GCN model for the prediction of Chemical Reactivity

Researchers

Connor W. Coley
MIT Research
Professor

Wengong Jin
Postdoctoral at
Broad Institute of
MIT and Harvard

Tommi Jaakkola
MIT CSAIL
Professor

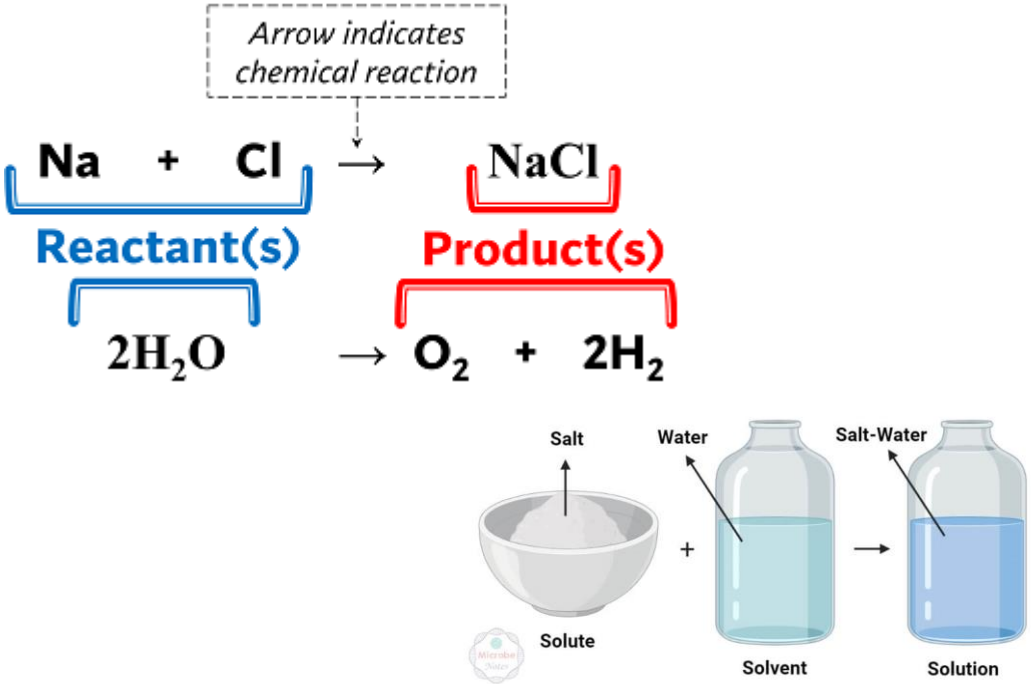
Klavs Jensen
MIT CE
Professor

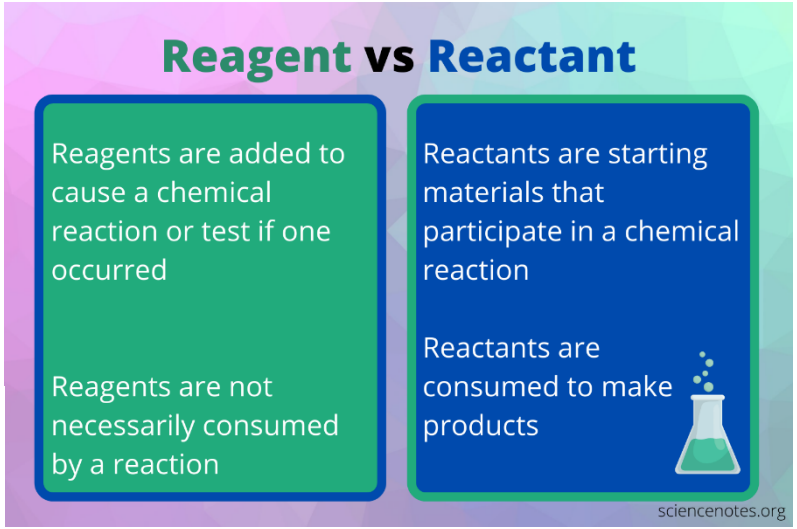
MT

1 Introduction

Main Points

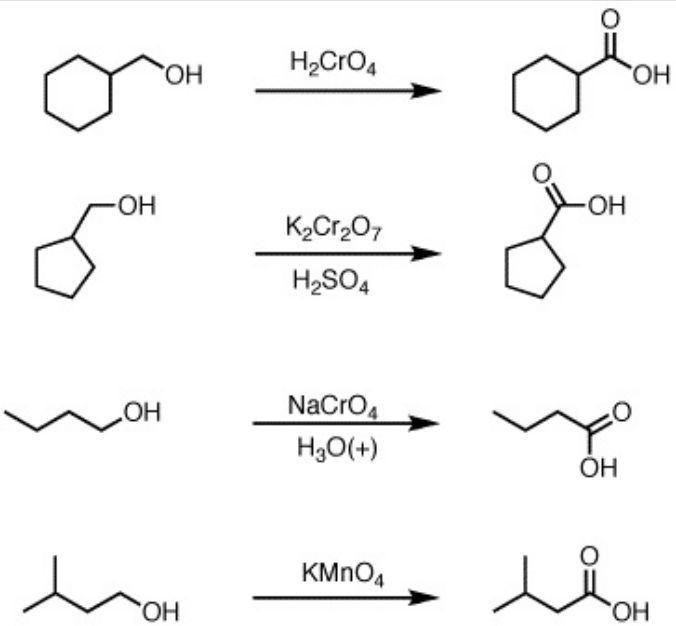
"Pioneer paper for a supervised learning approach to predict the products of organic reactions given their reactants, reagents, and solvents."

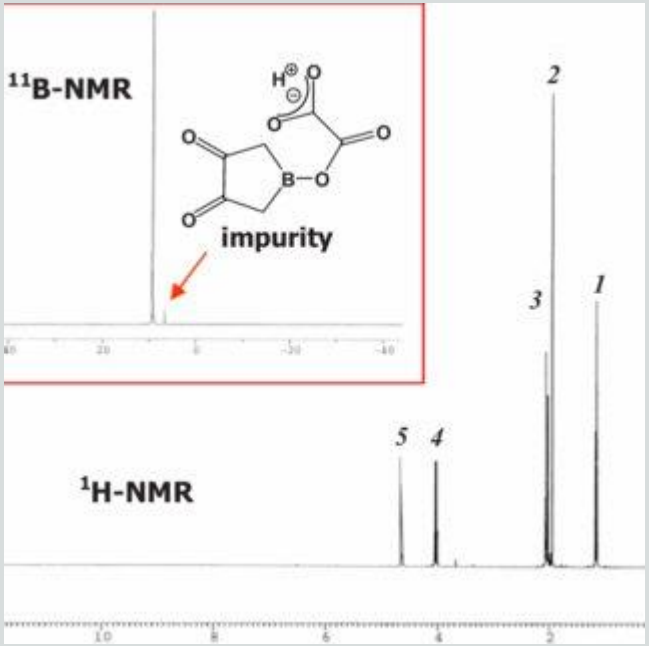




Goal

- Predict the major products of organic reactions based on reactant, reagent, and solvent
- Not just major products, but all species in a mixture of products
- Impurity Identification and quantification
- Reaction evaluation and drug substance manufacturing



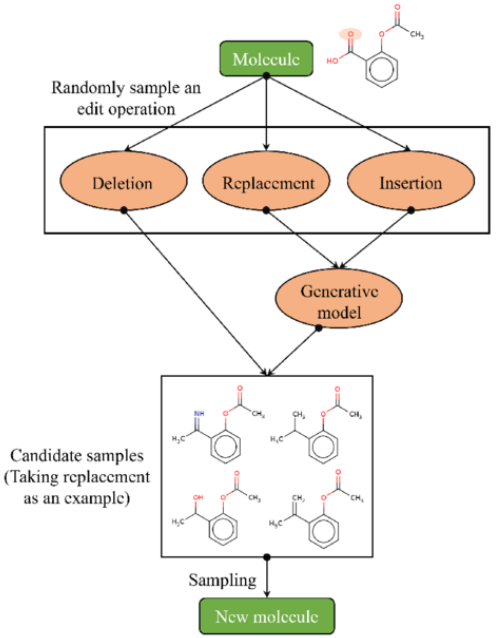


Prior Works

- **No Computational Methods**
  - Chemist imagine a sequence of possible chemical transformations
  - Then, perform the reactions in a laboratory, hoping that they happen as expected
  - But, sometimes chemical reactions often behave in unwanted ways
    - An undesired product is obtained instead of the desired one
    - The product yield is too low for the reaction to be useful
    - No reaction happens at all
- **Computer-Assistance in Chemical Synthesis**
  - CAMEO, EROS, IGOR, SOPHIA
    - expert heuristics to define possible mechanistic reactions
    - none achieved broad use within the chemistry community
  - Logic and Heuristics Applied to Synthetic Analysis (LHASA)
    - research group of Elias Corey at the Harvard university
    - uses AI techniques to discover sequences of reactions
    - combining the traditional use of reaction templates and ML
    - limitations on minor structural differences at the reaction center

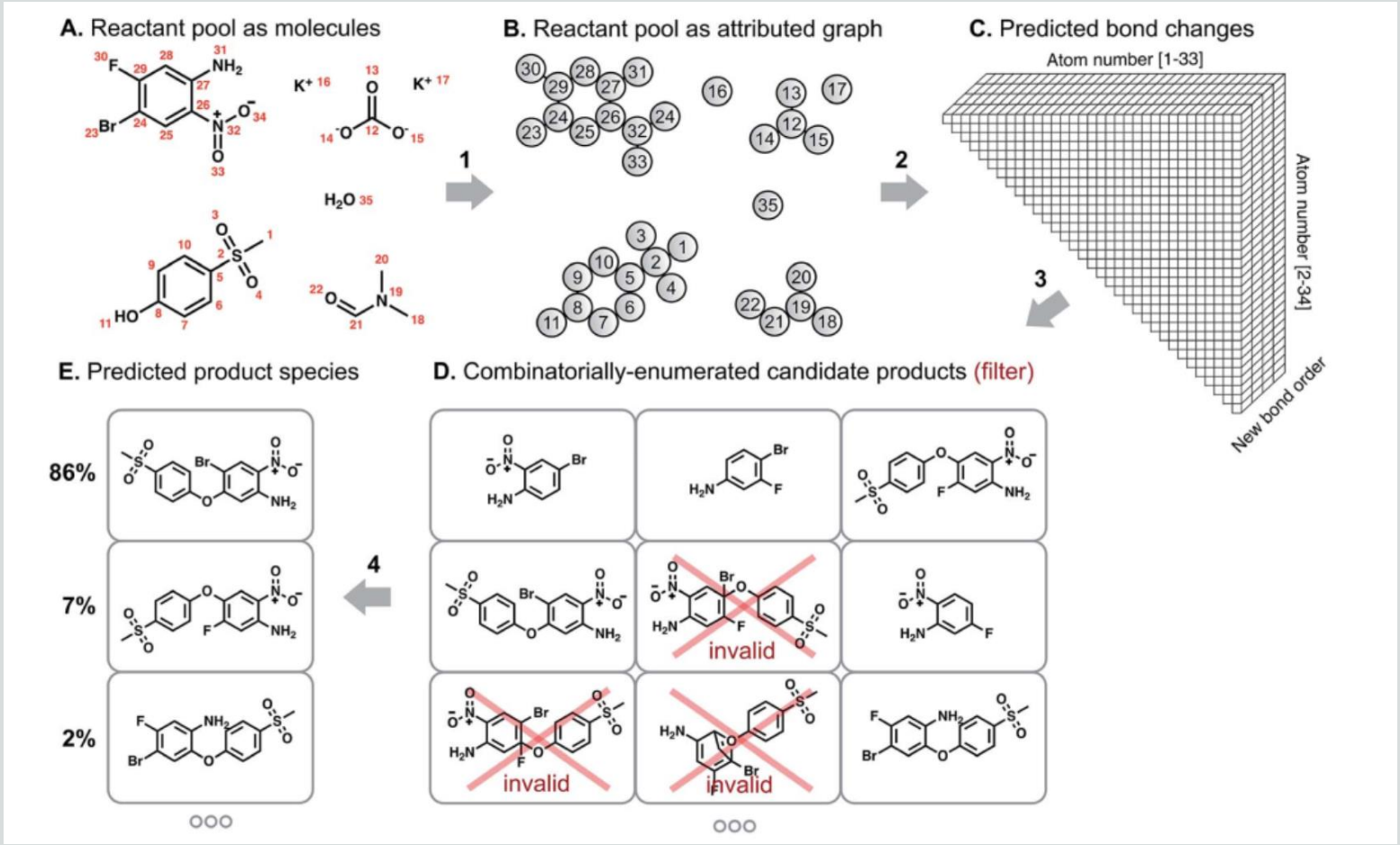
Main Points

- Graph representation of molecules: atom constitutes the nodes of the graph, while bond constitutes the edges of the graph.
- Predicting the outcome of a reaction is then equivalent to predicting graph edits i.e., which graph edges (representing chemical bonds) are changed



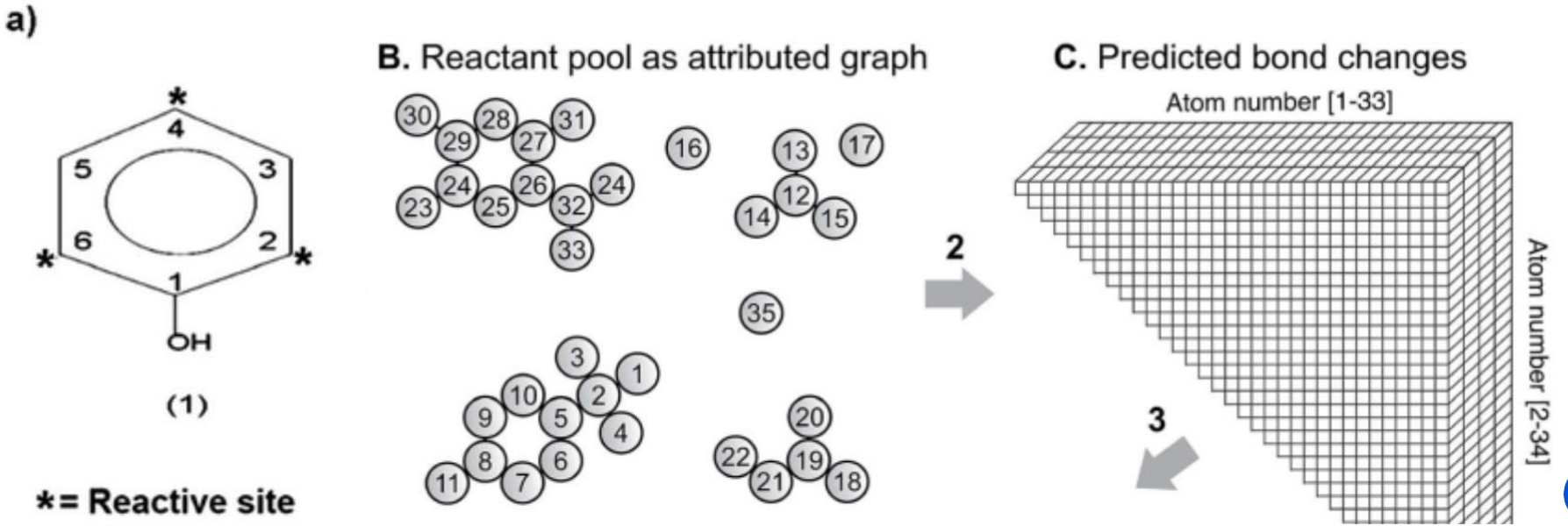
2 Methods

This Paper


A. Reactant pool as molecules
B. Reactant pool as attributed graph
C. Predicted bond changes
Atom number [1-33]
Atom number [2-34]
New bond order
1
2
3
4
H2O 35
E. Predicted product species
D. Combinatorially-enumerated candidate products (filter)
86%
7%
2%
invalid
invalid
invalid

Algorithm Process (1)

The algorithm mimics how chemist predict reaction outcomes
First, GCN tries to identify the reactive sites, i.e., atoms that are most likely to undergo a change in connectivity



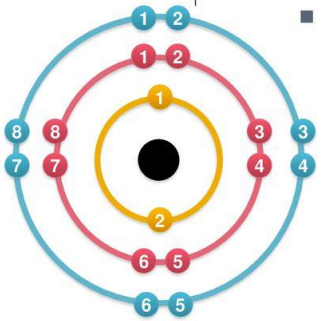
Algorithm Process (2)

Second, the products that could result from these change in chemical bonding are enumerated and filtered to eliminate molecule that don't satisfy chemical valence rules

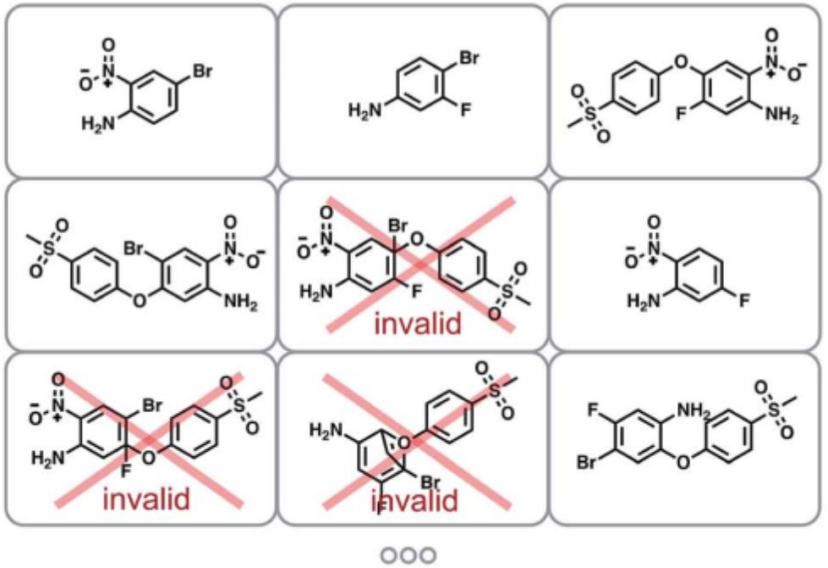
Valence Electrons

2-8-8 Rule

- Electrons orbit the nucleus in energy levels within the electron cloud
  - The 1st energy level holds 2 electrons
  - The 2nd and 3rd energy levels holds 8 electrons

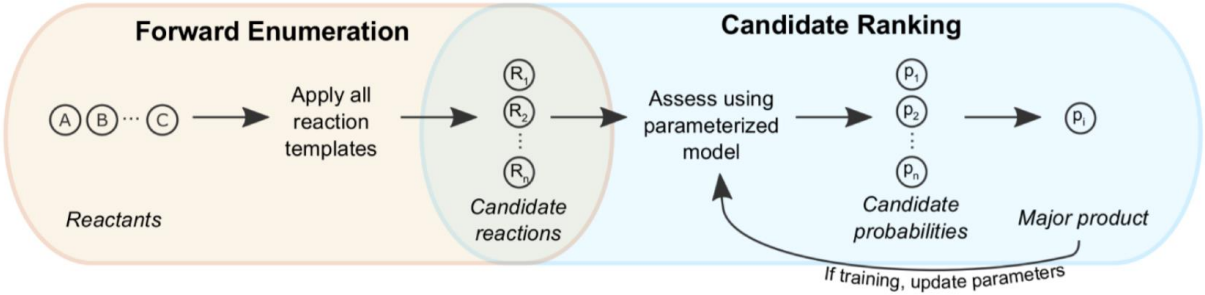


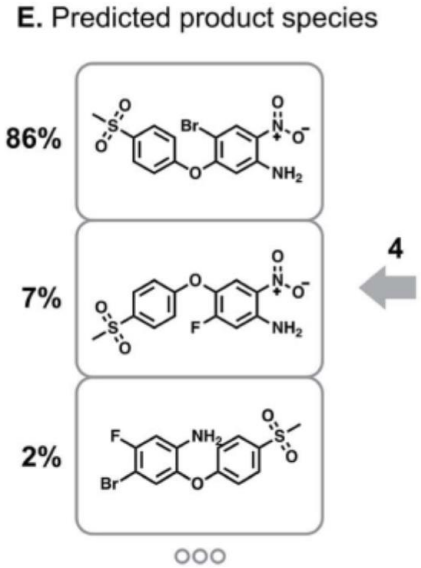
D. Combinatorially-enumerated candidate products (filter)



Algorithm Process (3)

Finally, another GCN scores the filtered molecules to yield a probability distribution over these possible reaction products

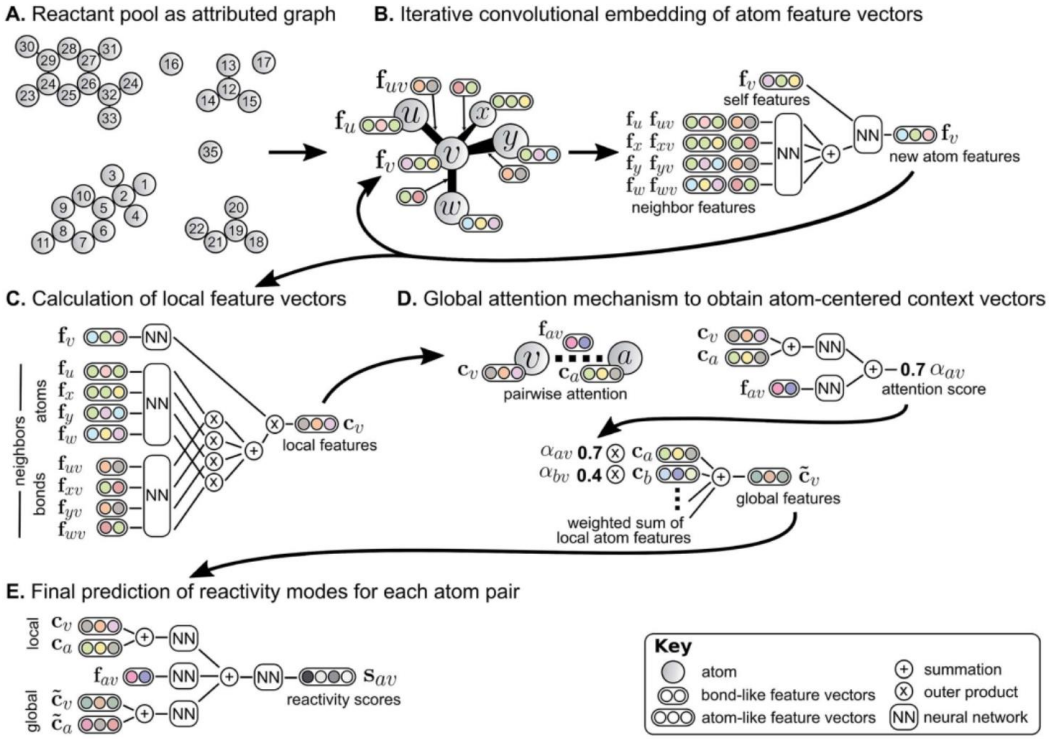




The primary aim of this work is the creation of the parameterized model, which is trained to maximize the probability assigned to the recorded experimental outcome

Prediction of Organic Reaction Outcomes Using Machine Learning
Connor et al., ACS Central Science 2017 (IF = 14.5)

Let's Dig Deeper

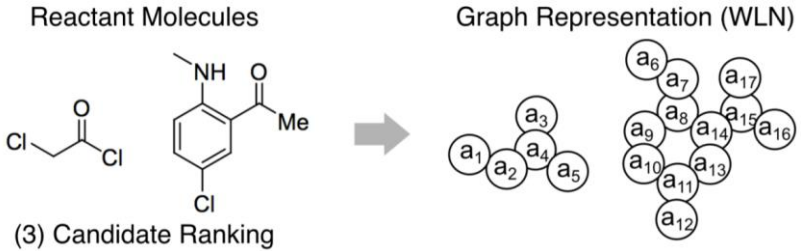


Algorithm Details (1)

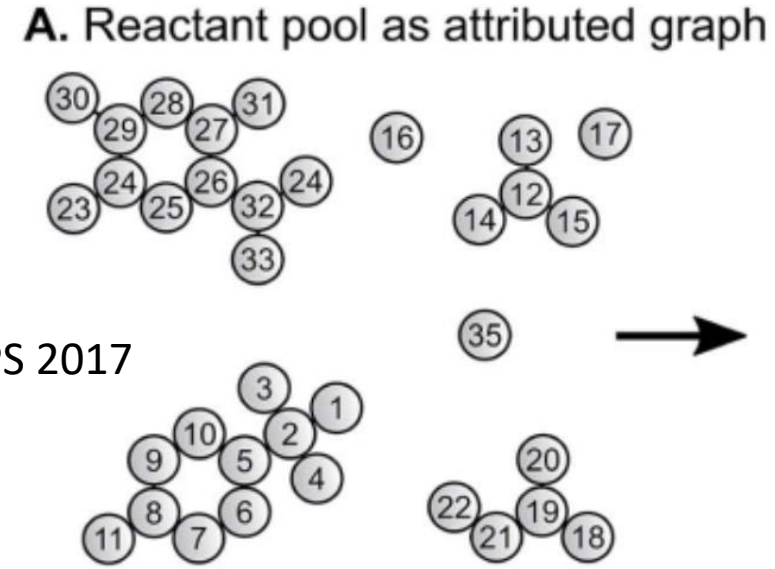
The authors used a Weisfeiler-Lehman Network (GCN model) First, graph representation of reactants is formed, where atoms are featured by atomic number, formal charge, etc… and bonds are featured by bond order and ring status. Atom-level features!!

Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network

Wengong Jin† Connor W. Coley‡ Regina Barzilay† Tommi Jaakkola†
†Computer Science and Artificial Intelligence Lab, MIT
‡Department of Chemical Engineering, MIT
†{wengong,regina,tommi}@csail.mit.edu, ‡ccoley@mit.edu

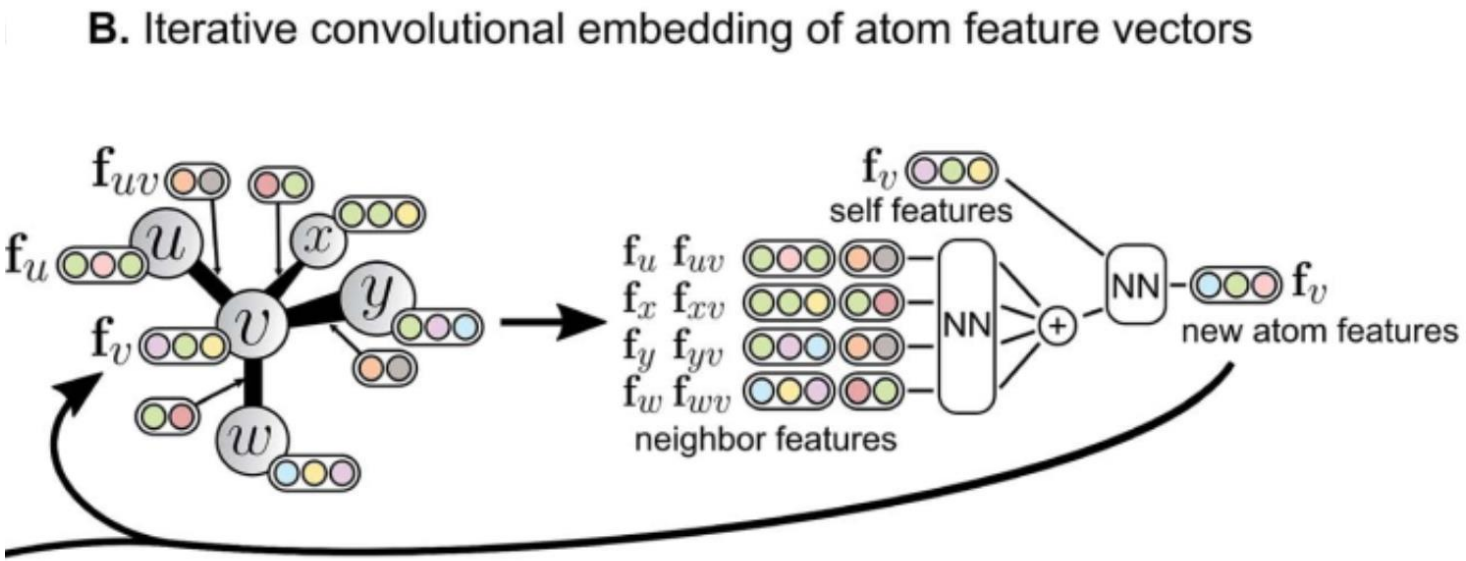


NIPS 2017



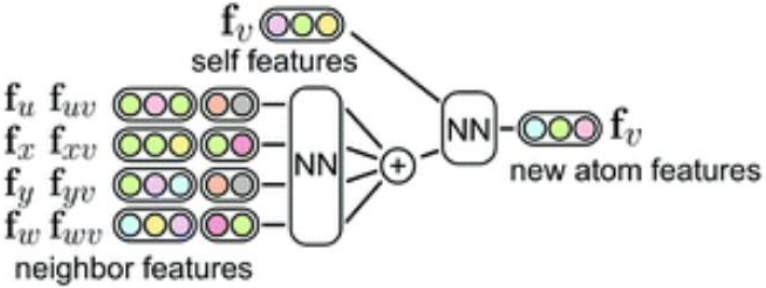
Algorithm Details (2)

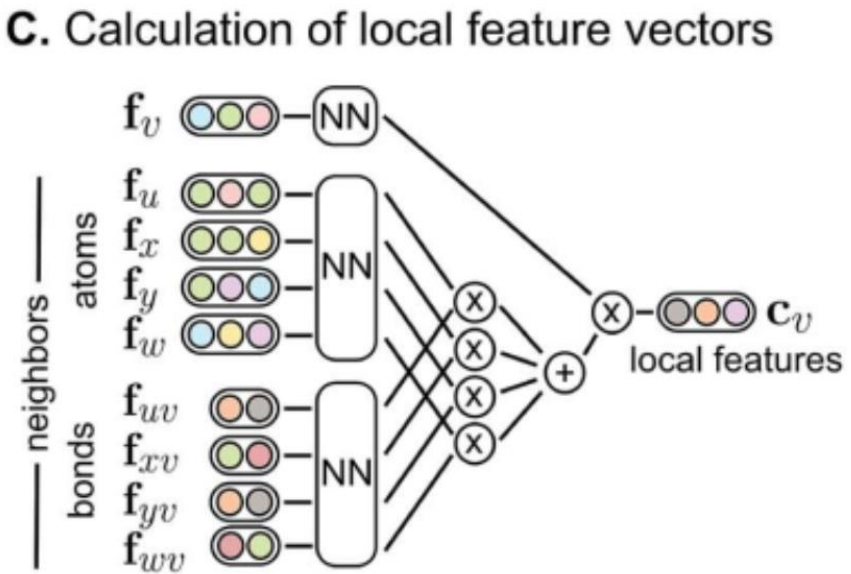
Second, atom-level features are iteratively updated by incorporating information from neighbor atoms.



Algorithm Details (3)

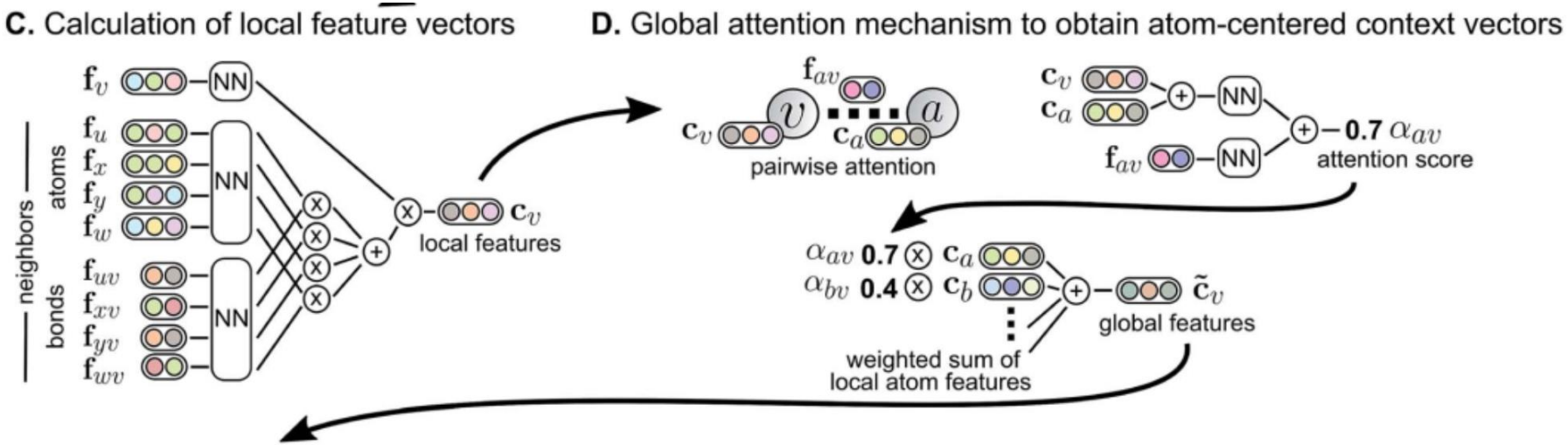
Third, the combination of atomic features obtained in the step B and bond features to obtain local feature vectors.



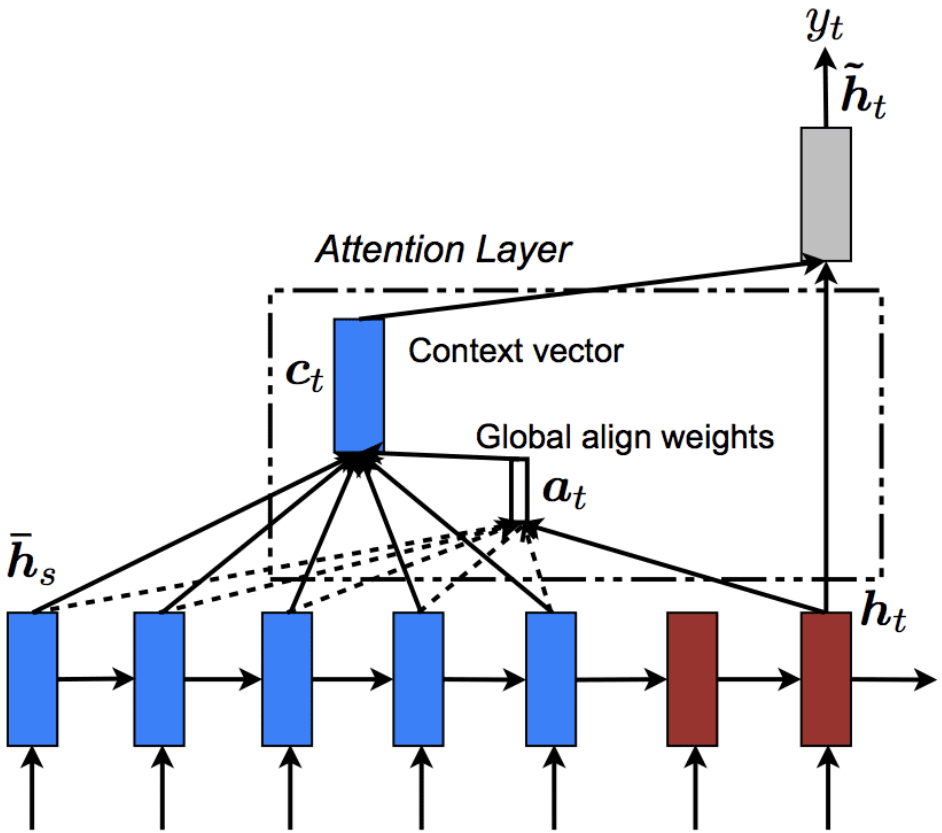


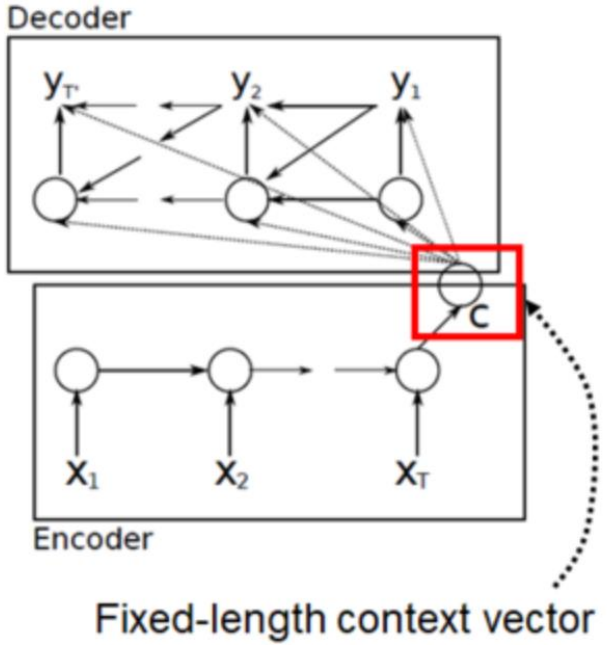
Algorithm Details (4)

Fourth, a global attention mechanism ensures that information about possible interactions with atoms that are distant are taken into account. The global attention mechanism produces atom-centered context vectors that take into account the effect of distant atoms



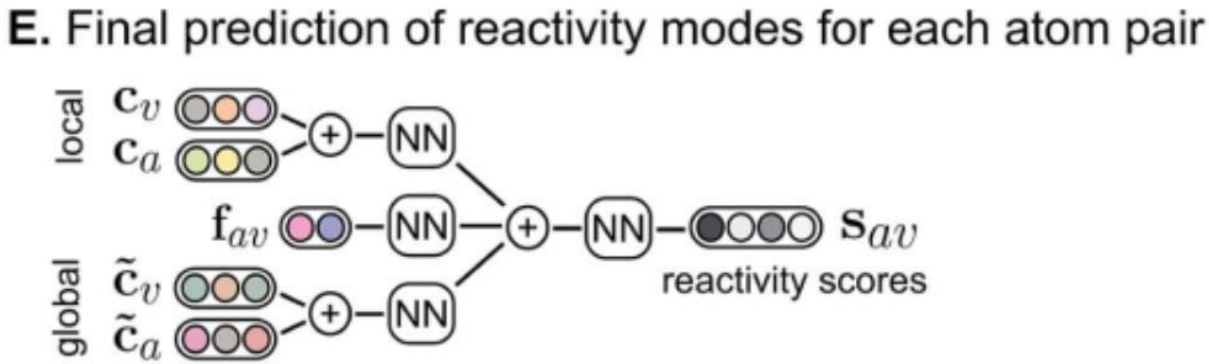
Recall


y_t
$\tilde{h}_t$
Attention Layer
c_t
Context vector
Global align weights
a_t
$\bar{h}_s$
h_t


Decoder
$y_{T'}$
y_2
y_1
C
x_1
x_2
x_T
Encoder
Fixed-length context vector

Algorithm Details (5)

Finally, the local and global context vectors are finally combined to obtain a probability of bond order change for each atom of the system



Check Points

- The probabilities of bond order changes are used to enumerate candidate products
- The authors restricted this enumeration to products that can be obtained by changing up to 5 unique bonds, and taking into account only the most likely bond changes.
- After this enumeration step, the obtained molecular structures were filtered using chemical valence rules to only keep molecules that can physically exist.
- The last part of the algorithm is the ranking of the candidate products. For this task, a second graph-convolutional neural network was used.
- Candidate outcomes produced by combination of more likely bond changes are themselves more likely to be the true outcome.

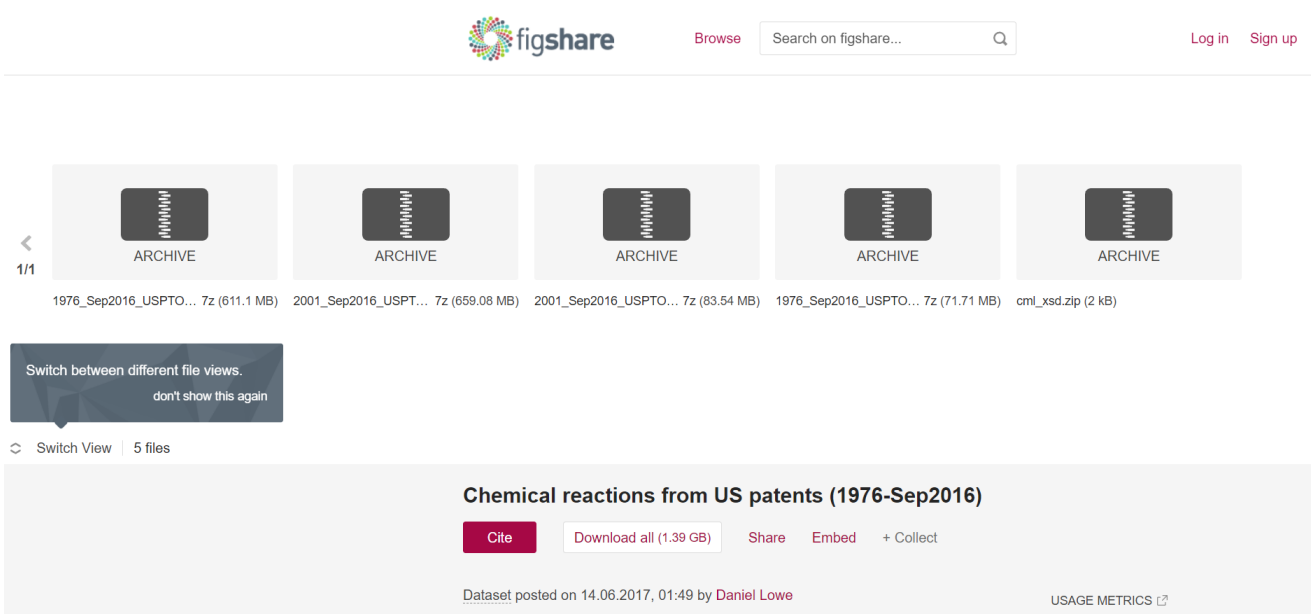
MT

3

Results

Dataset

The authors used a publicly available dataset, that was constituted from United States patents and comprises almost 410 000 reactions.



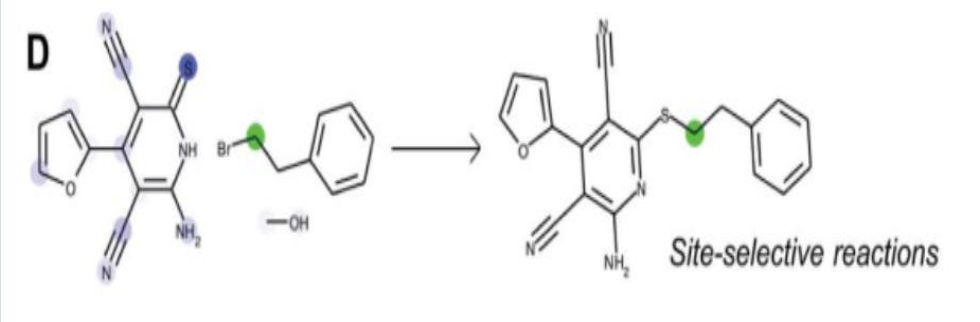
Result Table

For the algorithm presented in this article, the candidate with highest probability (Top-1) was the true reaction product in 85.6% of the cases, which is 5% higher than the previous best algorithm ("Sequence-to-sequence").

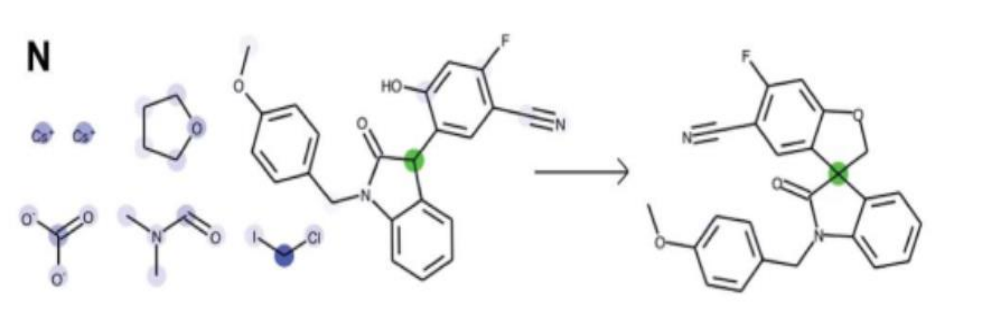
Method	$\|\theta\|$	Top-1 [%]	Top-2 [%]	Top-3 [%]	Top-5 [%]
WLN/WLDN[29]	3.2 M	79.6	—	87.7	89.2
Sequence-to-sequence[28]	30 M[a]	80.3	84.7	86.2	87.5
This work	2.6 M	**85.6**	**90.5**	**92.8**	**93.4**

[a] Estimated.

Human Benchmarking



The algorithm predicts the correct regioselectivity.



The complex formation of a quaternary carbon center is correctly predicted.

Summary

- **Combining two GCN model:** one for predicting the reactive sites and other for ranking the candidate products
- **Strength:** authors developed an algorithm that is able to reach a predictive power comparable to expert chemists.
- **Technical main contribution:** global attention mechanism
- **Limitation:** this algorithm does not take into account process parameters such as temperature, that can significantly change the outcome of a reaction.

Thank you