

Графовые нейронные сети для несвязанных графов в химических реакциях

Никитин Филипп

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов
Научный консультант PhD О. Исаев
Москва,
2020 г.



Требуется

Построить модель предсказания молекулярного графа основного продукта химической реакции по графам исходных веществ.

На модель накладываются ограничения:

- применима к данным в виде несвязанного молекулярного графа;
- допускает использования экспертных знаний о локальной структуре молекулярного графа;

Проблема

Пространство молекулярных структур высоко-размерное. Количество механизмов реакций растет с ростом числа известных структур.

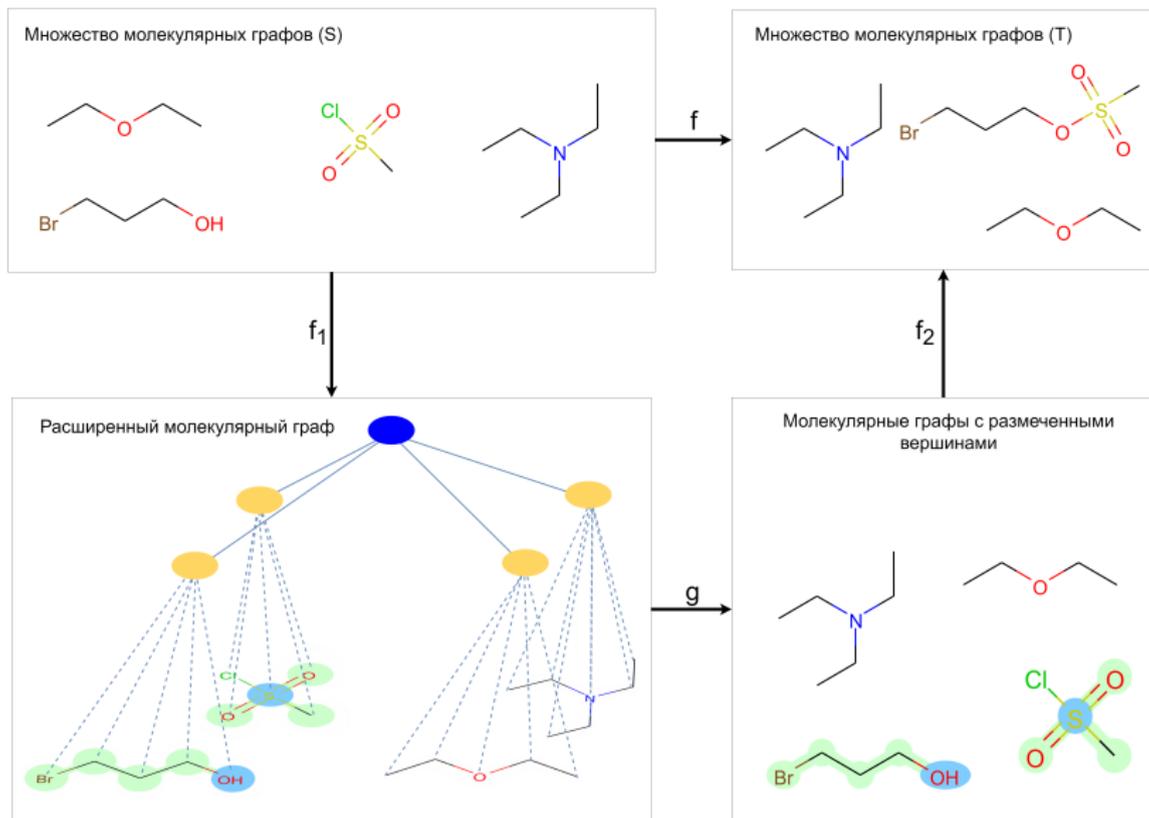
Метод

Графовая нейронная сеть, допускающая использование экспертных знаний о структуре молекулярного графа.

- 1 Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. *Machine learning for molecular and materials science*. Nature 2018, pp 547-559.
- 2 Schwaller P., Gaudin T., Lanyi D., Bekas C., Laino T. "Found in <https://www.overleaf.com/project/5ed3926f1393cb00011af764> Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models Chemical science 2018, 9, pp 6091–6098.
- 3 Jaworski W., Szymkuc S. *Automatic mapping of atoms across both simple and complex chemical reactions*. Nature communications 2019, 10, pp 1–11.
- 4 Corey E., Wipke W. T. *Computer-assisted design of complex organic syntheses*. Science 1969, 166, pp 178–192.

- 1 Schlichtkrull M., Kipf T. N., Welling M. *Modeling relational data with graph convolutional networks*. European Semantic Web Conference. 2018; pp 593–607.
- 2 Vaswani A., Shazeer N., Parmar N. *Attention is all you need*. Advances in neural information processing systems. 2017; pp 5998–6008.
- 3 Li J., Cai D., He X. *Learning graph-level representation for drug discovery*. arXiv preprint arXiv:1709.037412017.

Структура решения



Молекулярный граф

Атом

Атом — элемент конечного множества $a \in \mathcal{A} = \{a_1, a_2, \dots, a_n\}$, (C, N, S, Br).

Химическая связь

Химическая связь — элемент конечного множества $b \in \mathcal{B} = \{b_1, b_2, \dots, b_k\}$, (одинарная, двойная, водородная).

Молекула

Молекула — это планарный, неориентированный граф $M = (a, h, B)$, где:

- $a = [a_{k_1}, \dots, a_{k_l}]$ — упорядоченное множество атомов,
- $h = [h_1, \dots, h_l]$, где $h_i \in \mathbb{H}$, пространство описаний атомов,
- B — матрица смежности $l \times l$, $b_{i,j} \in \mathcal{B}$ — тип связи между a_{k_i} и a_{k_j} .

Химическая реакция — отображение исходных веществ в продукты

Химическая реакция

Дано:

- исходные вещества — множество $S = \{M_1, \dots, M_m\}$,
- продукты — множество $T = \{M_1, \dots, M_k\}$.

Химическая реакция — отображение $f : S \rightarrow T$, где $f \in \mathcal{F}$.

Задача выбора модели

Задано семейство параметрических функций \mathcal{F} (в работе графовые CNN)

$$f = \underset{f \in \mathcal{F}}{\operatorname{arg\,min}} L(T, f(S)),$$

где L целевая функция потерь.

База реакций

База реакций из американских патентов США, Lowe, 2012:

- 1 млн. реакций в формате SMARTS,
- 1976-2016 год регистрации патента,
- Разделены катализаторы и реагенты,
- Для заданных реагентов, катализаторов известен основной продукт.

SMILES — язык, который позволяет однозначно закодировать молекулярный граф строкой символов ASCII, SMARTS — надстройка SMILES, позволяющая специфицировать меж-структурные взаимосвязи в молекулах.

Пример: c1cccc[c:1]1[NH2:2]>>c1cccc[c:1]1[N:2](=O)=O

Основной продукт — молекула, включающая в себя наибольшее количество атомов среди продуктов реакции.

RDKit — библиотека для работы с молекулярными графами.

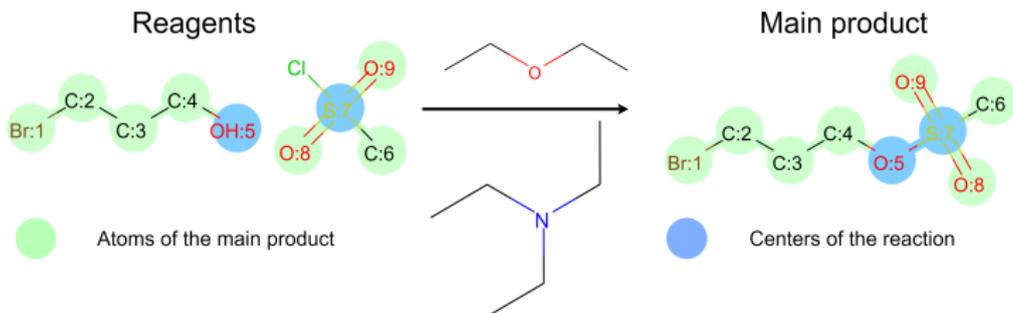
Классификация вершин в графе

Определение атомов основного продукта

Для множества атомов исходных веществ требуется определить вероятность принадлежности к множеству атомов основного продукта.

Определение центров реакции

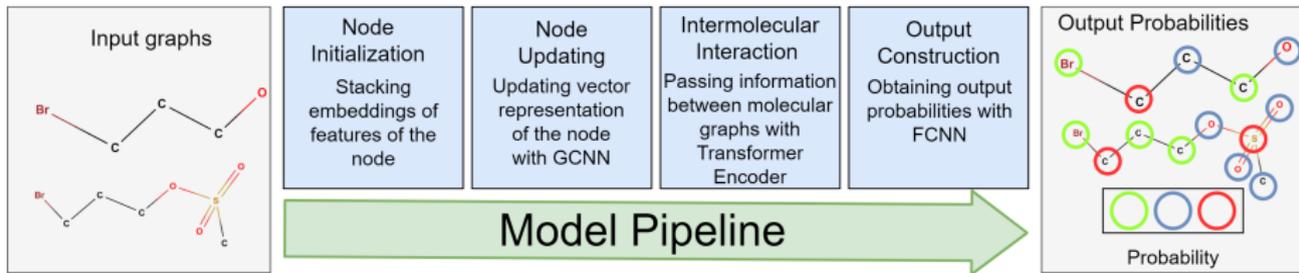
Для множества атомов основного продукта требуется выделить те, конфигурация которых изменилась в течение реакции.



Химическая реакция отображает реагенты в продукты. Выделены атомы основного продукта и центры реакции.

Модель классификации вершин в несвязанном графе

- Эмбединг признаков вершин молекулярного графа.
- Обновление состояний вершин графовой сверточной сетью.
- Преобразование состояний вершин кодировщиком архитектуры Transformer.
- Получение вероятностей меток атомов полносвязанной нейронной сетью.



Формула обновления состояний вершин

$$h_i^{(l+1)} = \sigma \left(W^{(l)} h_i^{(l)} + \sum_{r \in R} \sum_{j \in N_i} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} \right),$$

R — множество типов ребер графа, типов химической связи,

W, W_r — параметры преобразования,

$h_i^{(l)}$ — векторное представление вершины графа, атома a_i в слое l ,

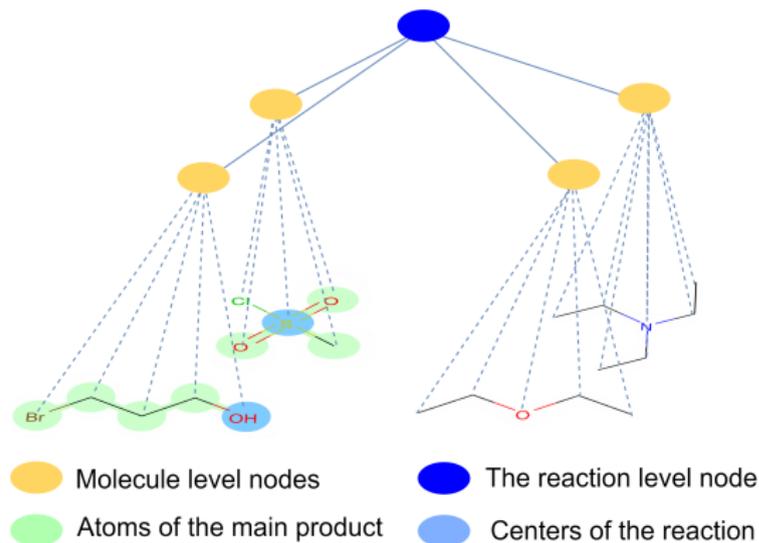
$c_{i,r}$ — нормировочный множитель, обычно кратность вершины графа,

σ — нелинейная функция

Проблема

$h_i^{(l+1)}$ зависит только от векторных представлений вершин $h_j^{(l)}$ той же компоненты связности, что и a_i . Химическая реакция обусловлена межмолекулярным взаимодействием.

Расширенный граф химической реакции



Расширенный граф химической реакции: введены вершины, соответствующие векторным описаниям молекул и всей химической реакции.

Обновление векторных состояний вершин

$$h_i^{(l+1)} = \text{ReLU} \left(W^{(l)} h_i^{(l)} + W_{ml}^{(l)} h_{m_k}^{(l)} + \sum_{r \in R} \sum_{j \in N_i} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} \right),$$

$$h_{m_k}^{(l+1)} = \text{ReLU} \left(W^{(l)} h_{m_k}^{(l)} + W_{rl}^{(l)} h_r^{(l)} + \sum_{j \in m_k} \frac{1}{|m_k|} W_{ml}^{(l)} h_j^{(l)} \right),$$

$$h_r^{(l+1)} = \text{ReLU} \left(W^{(l)} h_r^{(l)} + \sum_{m_j \in M} \frac{1}{|M|} W_{rl}^{(l)} h_{m_j}^{(l)} \right).$$

$h_i^{(l+1)}$ — векторное представление атома

$h_{m_k}^{(l+1)}$ — векторное представление молекулы

$h_r^{(l+1)}$ — векторное представление реакции

Self-Attention

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{KQ^T}{\sqrt{d_{\text{model}}}} \right) V.$$

Transformer

$$\begin{aligned} H_{\text{mha}}^{(l)} &= \text{concat}[\text{head}_1, \text{head}_2, \dots, \text{head}_h] W^O, \\ \text{head}_i &= \text{Attention} \left(H^{(l)} W_i^Q, H^{(l)} W_i^K, H^{(l)} W_i^V \right). \end{aligned}$$

Векторные состояния в матрице $H_{\text{mha}}^{(l)}$ есть выпуклые комбинации векторных состояний из матрицы $H^{(l)}$ с оптимизируемыми коэффициентами.

Эволюционное семейство моделей:

- 1 Базовая модель (BASE) не учитывает типы ребер, признаки атомов; не использует механизмы работы с несвязанными графами.
- 2 Модель расширенного молекулярного графа (EG). По сравнению с базовой моделью, использует расширенный молекулярный граф.
- 3 Модель Трансформер (T). По сравнению с базовой моделью, после сверточных слоев используется преобразование self-attention.
- 4 Модель EGT Использует обе предложенных модификации.
- 5 Модель EGTB использует разные типы ребер в соответствии с типом химической связи.
- 6 Модель EGTBF использует признаки атомов (валентность, заряд и тд).
- 7 Модель MT_EGTBF использует многозадачное обучение для двух рассматриваемых задач.

Сводная таблица результатов

	Product mapping		Center detection	
	FM	F_1	FM	F_1
BASE	0.21 ± 0.01	0.92 ± 0.002	0.15 ± 0.01	0.502 ± 0.002
EG	0.45 ± 0.01	0.943 ± 0.002	0.40 ± 0.01	0.714 ± 0.002
T	0.36 ± 0.01	0.938 ± 0.002	0.29 ± 0.01	0.643 ± 0.002
EGT	0.47 ± 0.01	0.946 ± 0.002	0.43 ± 0.01	0.731 ± 0.002
EGTB	0.53 ± 0.01	0.950 ± 0.002	0.55 ± 0.01	0.809 ± 0.002
EGTBF	0.59 ± 0.01	0.959 ± 0.002	0.60 ± 0.01	0.838 ± 0.002
MT_EGTBF	0.60 ± 0.01	0.963 ± 0.002	0.61 ± 0.01	0.841 ± 0.002

FM среднее значение точности полного совпадения (1, если все метки атомов в реакции предсказаны верно, 0 иначе). F_1 среднее значение F_1 -меры между предсказанными и правильными метками атомов в реакции.

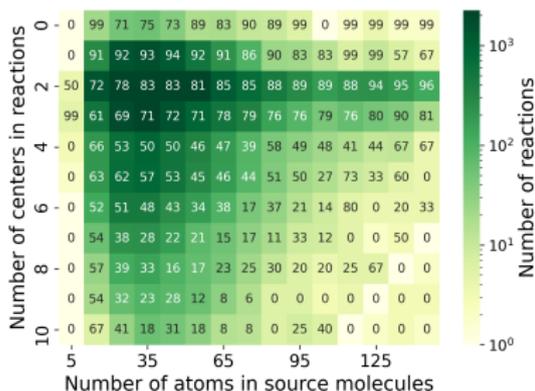
Вывод

Предложенные методы работы с несвязанными графами значительно улучшают качество модели. Использование признаков вершин и ребер молекулярного графа приводит к повышению качества.

Анализ ошибки

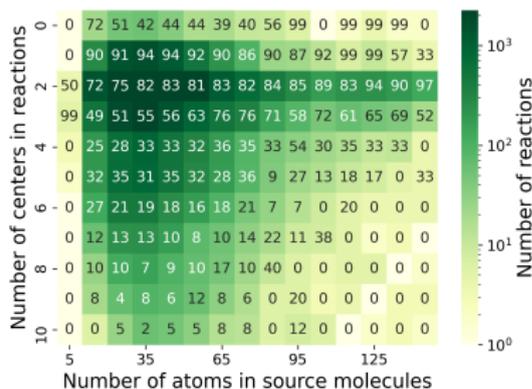
На графике представлена совместная зависимость качества модели от количества центров и длины исходных молекул. Цветом указано распределение исходных данных.

Detection of atoms of the main product



50 The number means full-match accuracy

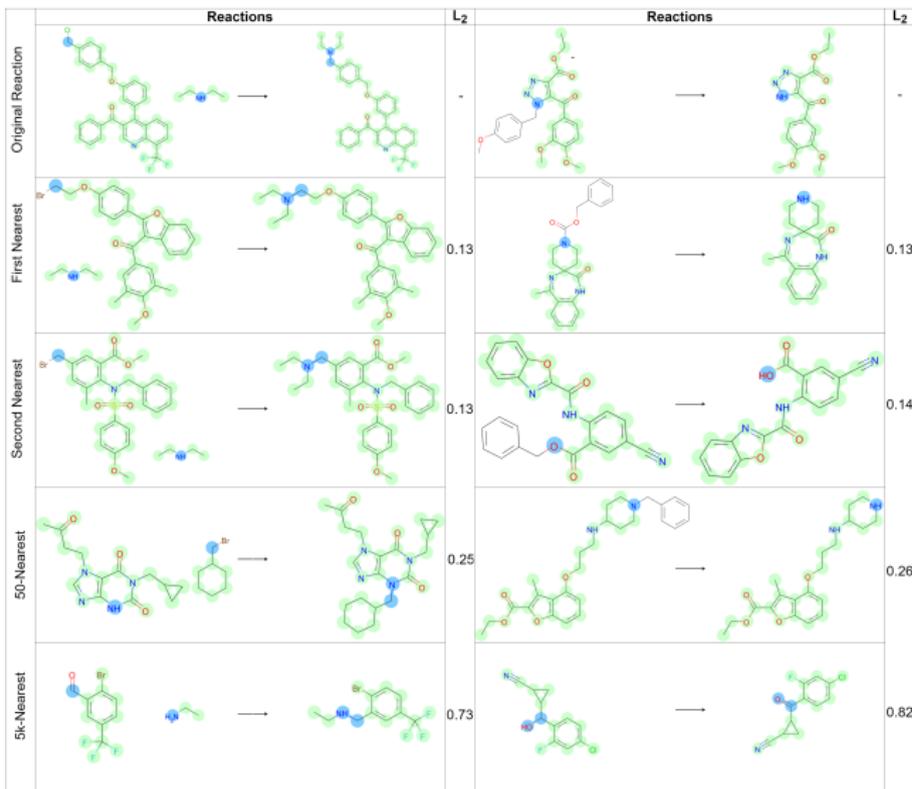
Detection of centers of the reaction



The color means the number of reactions

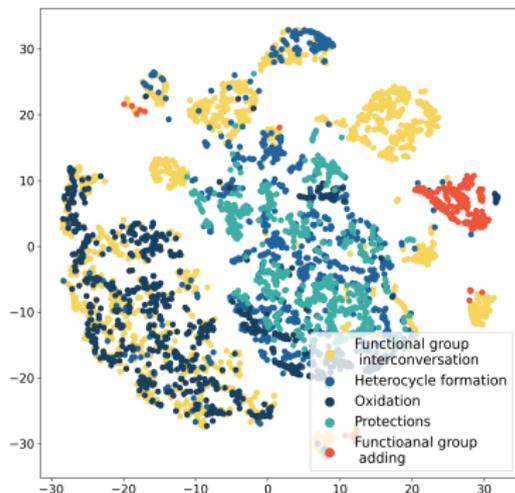
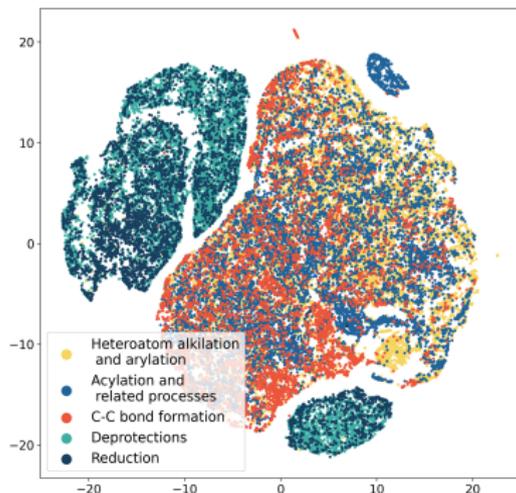
- 1 С увеличением числа центров качество падает;
- 2 Зависимость качества от длины исходных молекул выражена слабо.

Исследование свойств пространства реакций



Метрически близким векторам, соответствующим состояниям всей реакции, соответствуют реакции с похожим механизмом.

T-SNE карты реакций



Для датасета USPTO_50k (10 классов реакций) построены T-SNE проекции в двумерное пространство.

Кластеры векторов состояний реакции скоррелированы с разметкой по классам химических реакций.

- 1 Сформулирована задача предсказания продуктов химической реакции в терминах классификации вершин несвязанного графа
- 2 Предложено обобщение графовых нейронных сетей для работы с несвязанными графами
- 3 Предложена последовательность вычислительных экспериментов, демонстрирующая необходимость каждой предложенной модификации
- 4 Проанализирована полученная модель, исследованы свойства векторных состояний химической реакции, формируемых в модели

 **Filipp Nikitin, Olexandr Isayev, and Vadim Strijov.**

Dracon: Disconnected graph neural network for atom mapping in chemical reactions.

Physical Chemistry Chemical Physics, 2020.