Построение графовых нейронных сетей в задаче синтеза химических молекул

Авторы: Никитин Ф., Стрижов В.

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(государственный университет)
Факультет управления и прикладной математики

27 ноября 2019 г.

Графовые нейронные сети в задаче синтеза молекул

Актуальность

Открытие молекул с заданными свойствами занимает до **10 лет** и стоит в среднем **7 млрд. долларов**. Синтез вещества с заданной молекулярной структурой — решаемая в процессе задача.

Требуется

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

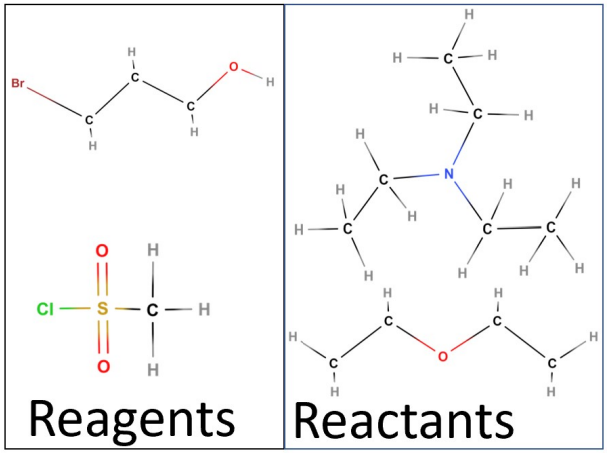
Проблема

Задача не решена на приемлимом уровне качества. Существующие решения не имеют явного потенциала для равития.

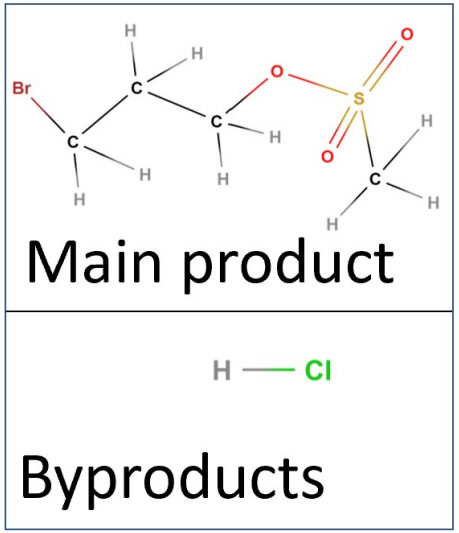
Метод

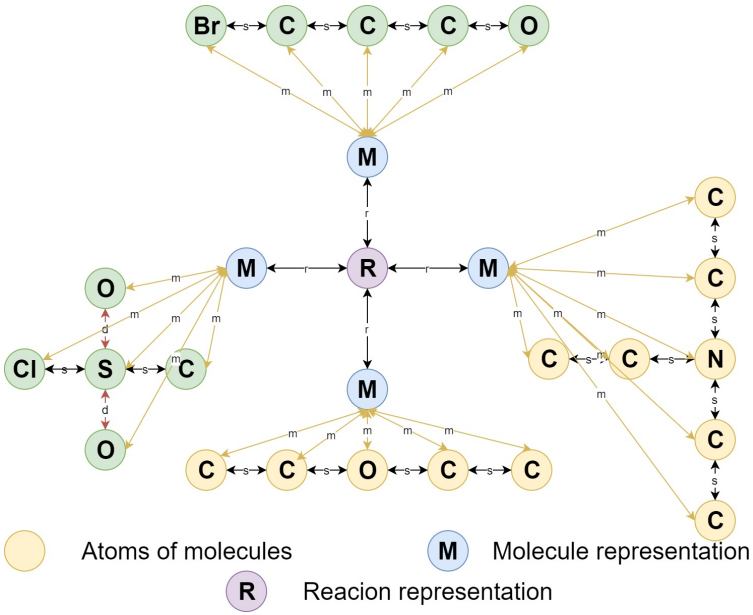
Графовая нейронная сеть, допускающая использование экспертных заний о структуре молекулярного графа.

Структура решения

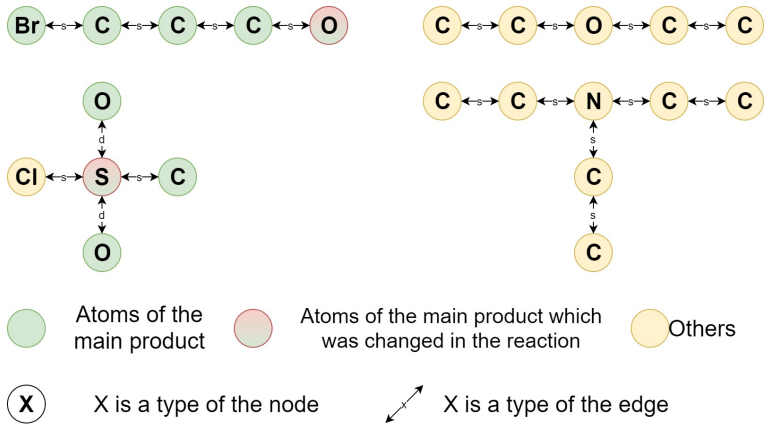


f





g



Обзор литературы

SMILES2SMILES translation, Schwaller, 2018

- Демонстрируют адекватные результаты
- Алгоритмы активно развиваются
- Не учитывается грамматика языка SMILES
- Не используются химические свойства элементов молекулярного графа
- Не очевиден потенциал для улучшения
- Избыточное количество параметров

Основанные на правилах ассистенты: SOFIA, IGOR, CAMEO

- Основаны на экспертных знаниях
- Интерпретируемы
- Не имеют обобщающей способности
- С развитием органической химии количество правил увеличивается
- Демонстрируют неудоволетворительные результаты

Молекулярный граф

Атом

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

Химическая связь

Химическая связь — элемент конечного множества $b \in \mathcal{B} = \{b_1, b_2, \ldots, b_k\}$ (одинарная, двойная, водородная), взаимодействие атомов, обусловливающее устойчивость молекулы или кристалла как целого.

Молекула

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

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

Химическая реакция — отображение

Химическая реакция

Дано:

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

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

Задача выбора модели

Задано семейство параметрических функций $\mathcal{F}$

$$f = arg \min_{f \in \mathcal{F}} \mathbb{E}_{S,T} L(T, f(S))$$

где L целевая функция потерь.

Формат исходных данных

База реакций

База реакций из американских патентов США, Lowe, 2012

1. 3 млн. реакций в формате SMARTS
2. 1976-2016 год регистрации патента
3. Разделены катализаторы и реагенты
4. Для заданных реагентов, катализаторов известен основной продукт

Примечания

SMILES — язык, который позволяет однозначно закодировать молекулярный граф строкой символов ASCII, SMARTS — надстройка SMILES, позволяющий специфицировать межструктурные взаимосвязи в молекулах.
Пример: $\mathrm{c1ccccc[c:1]1[NH2:2]\gg c1ccccc[c:1]1[N:2](=O)=O}$
Основной продукт — молекула, включающая в себя наибольшее количество атомов среди продуктов реакции.
RDKit — библиотека, позволяющая преобразовать исходные SMARTS в молкулярные графы, определить признаки связей.

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 течении реакции.

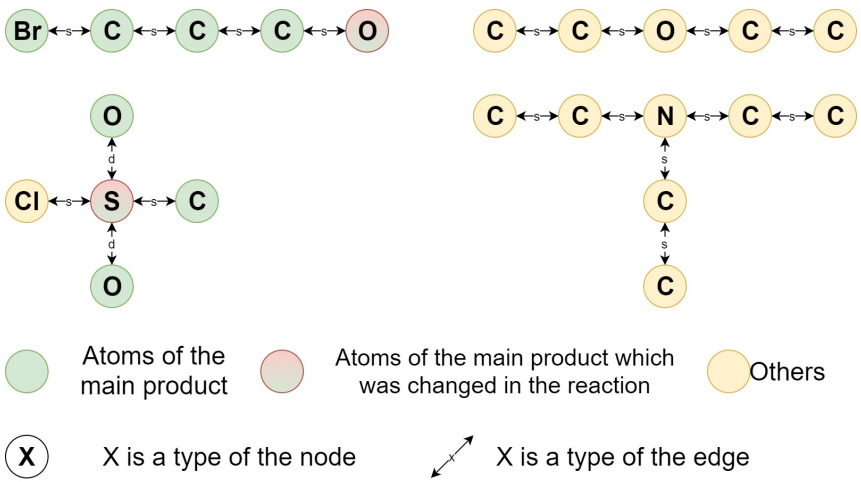


Рис.: Молекулярные графы исходных веществ с размеченными классами атомов

Архитектура модели

- Инициализация векторных представлений вершин
- Графовая свёрточная нейронная сеть
- Кодировщик архитектуры Transformer
- Полносвязанная нейронная сеть для классиификации векторных представлений вершин графа

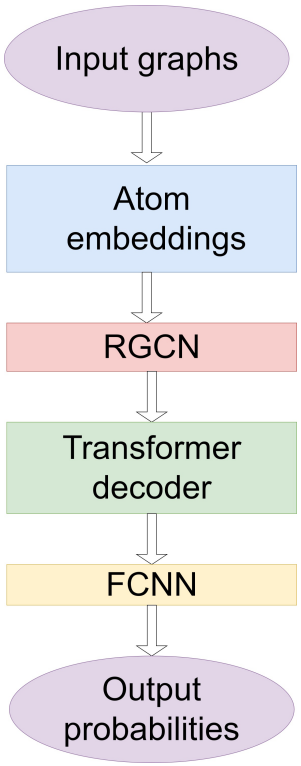


Рис.: Архитектура модели

Relational Graph Convolution Neural Network

RGCNN, Schlichtkrull, 2018

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right) \quad (1)$$

R — множество типов ребер графа, типов химической связи,
$\mathbf{W}, \mathbf{W}_r$ — параметры преобразования,
$\mathbf{h}_i^{(l)}$ — вкторное представление вершины графа, атома a_i в слое l,
$c_{i,r}$ — нормировочный множитель, обычно кратность вершины графа
σ — нелинейная функция

Проблема

$\mathbf{h}_i^{(l+1)}$ зависит только от векторных представлений вершин $\mathbf{h}_j^{(l)}$ той же компоненты связанности, что и a_i. Химическая реакция обусловлена внешмолекулярным взаимодействием.

Расширенный граф химической реакции

Обновление векторных состояний вершин

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}_{ml}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right) \tag{2}$$

$$\mathbf{h}_{m_k}^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}_{rl}^{(l)}\mathbf{h}_r^{(l)} + \sum_{r\in R}\sum_{j\in M_k}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right) \tag{3}$$

$$\mathbf{h}_r^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_r^{(l)} + \mathbf{W}_{rl}^{(l)}\mathbf{h}_r^{(l)} + \sum_{j\in M}\frac{1}{|M|}\mathbf{W}_{rl}^{(l)}\mathbf{h}_{m_j}^{(l)}\right) \tag{4}$$

$\mathbf{h}_i^{(l+1)}$ — векторное представление атома
$\mathbf{h}_{m_k}^{(l+1)}$ — векторное представление молекулы
$\mathbf{h}_r^{(l+1)}$ — векторное представление реации

Расширенный граф химической реакции

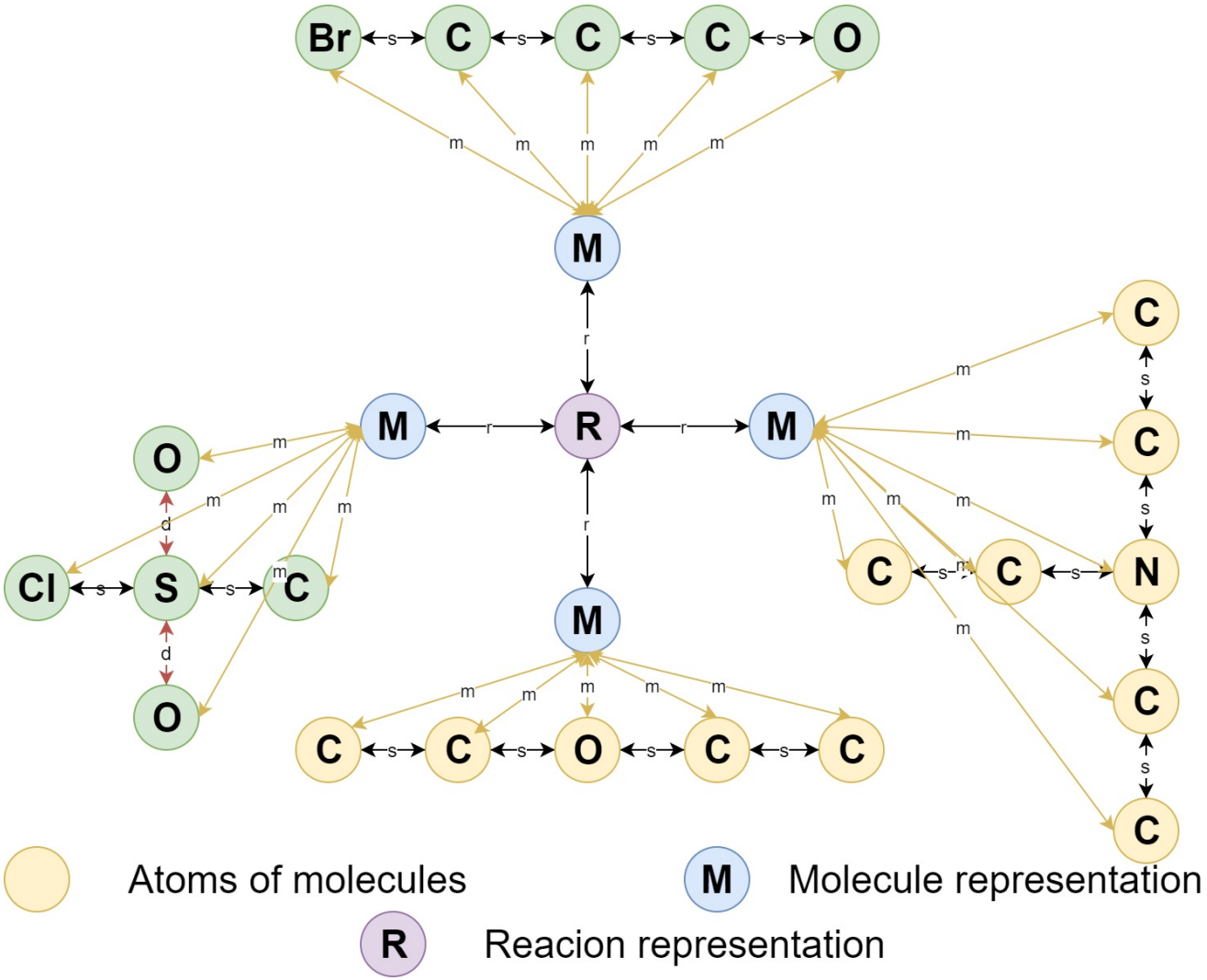


Рис.: Расширенный граф химической реакции: введены вершины соответствующие векторным описаниям молекул и всей химической реакции

Моделирование взаимодействия атомов разных молекул

Transformer, Vaswani, 2017

$$\mathbf{H}^{(l)} = [\mathbf{h}_1^{(l)}, \ldots, \mathbf{h}_n^{(l)}] \tag{5}$$

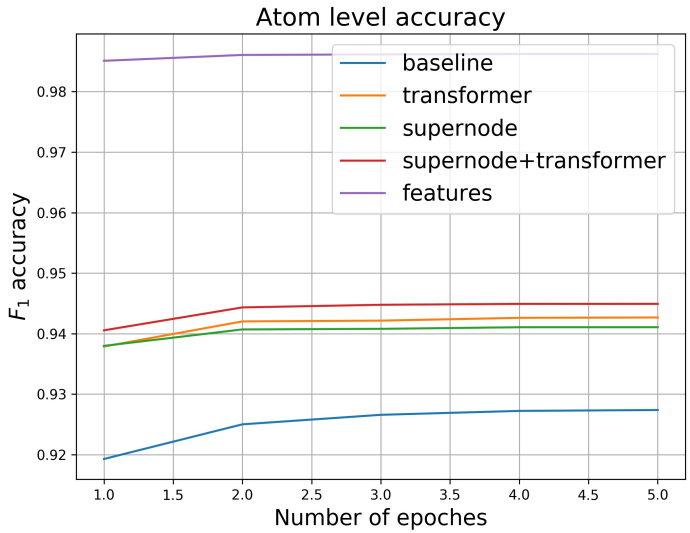
$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{K}\mathbf{Q}^\top}{\sqrt{d_{model}}})\mathbf{V} \tag{6}$$

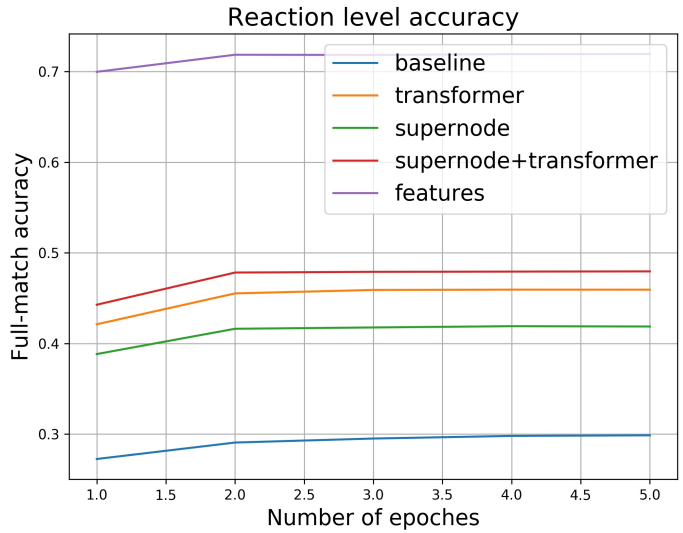
$$\mathbf{H}_{mha}^{(l)} = [head_1, head_2, \ldots, head_h]\mathbf{W}_O \tag{7}$$

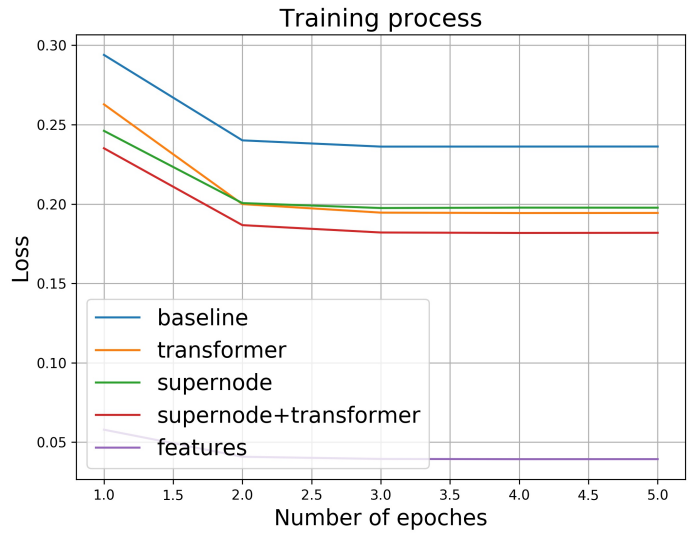
$$head_i = Attention(\mathbf{H}^{(l)}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{H}^{(l)}\mathbf{W}_i^{\mathbf{K}}, \mathbf{H}^{(l)}\mathbf{W}_i^{V}) \tag{8}$$

- Векторные состояния в матрице $\mathbf{H}_{mha}^{(l)}$ есть выпуклые комбинации векторных состояний из матрицы $\mathbf{H}^{(l)}$ с оптиизируемыми коэфициентами
- Преобралование также включает последовательность линейых преобразований с нелинейностями

Качество выделения атомов основного продукта

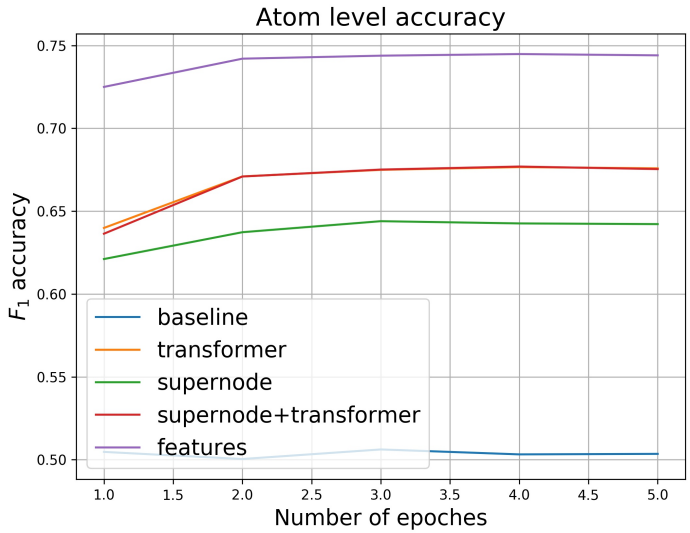


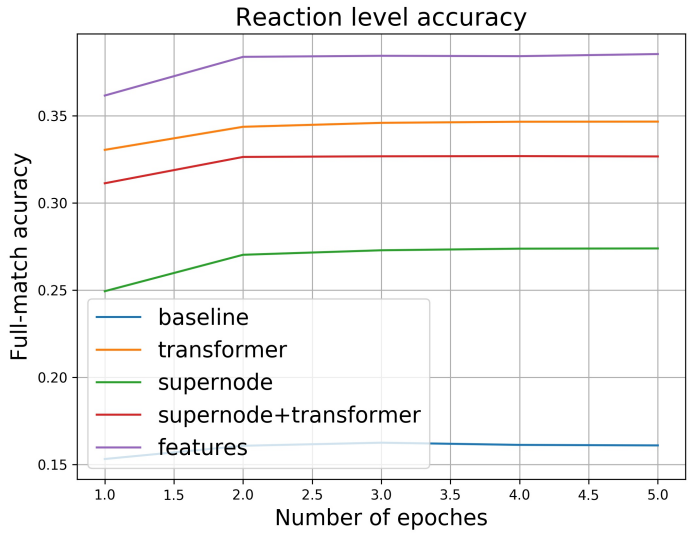


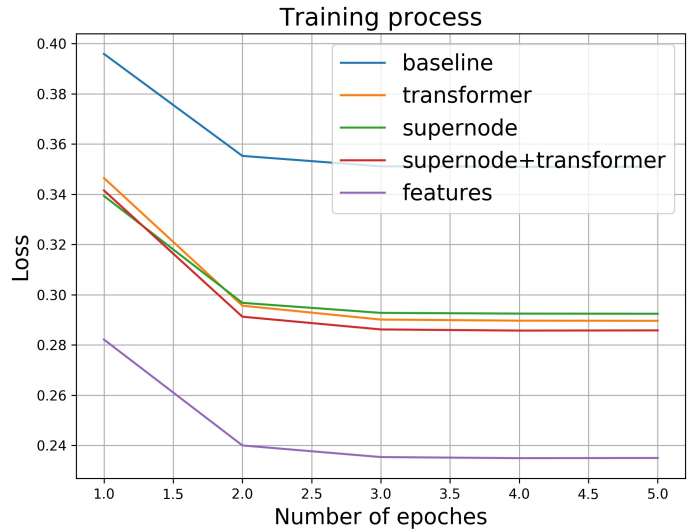


- Модель демонстрирует адекатные результаты
- Использоание знаний о структуре молекулярного рафа приводит к росту качества
- лучшая модел для 7 из 10 реакций точно определяет все атомы продукта

Качество выделения центров реакции







- Модель демонстрирует адекатные результаты
- Использоание знаний о структуре молекулярного рафа приводит к росту качества
- лучшая модел для 5 из 10 реакций точно определяет все центры

Сводная таблица резльтатов

	Product mapping		**Center detection**	
	FM	F_1	*FM*	F_1
baseline	0.29	0.929	0.16	0.502
supernode	0.42	0.942	0.28	0.641
transformer	0.46	0.944	0.35	0.677
transformer + supernode	0.48	0.946	0.33	0.672
features	0.73	0.985	0.45	0.75

FM — точность определения всех атомов в рекции

F_1 — среднее значене F_1меры между предсказанными и оригиальными метками атомов реакции

Текущие результаты

Выводы

1. Предложена новая архитектура модели предсказания основного продукта химической реакции
2. Экспериментально показана адекватность модели
3. Предложены различные способы работы нейронных сетей для несвязанных графов
4. Показано, что внесение информации о структуре молекулярного графа и химических свойствах атомов приводит к улучшению результатов

Планируемые исследования

1. Провести анализ параметров архитектуры, выявить закономерности согласующиеся с физикой процесса
2. Построить решение задачи ретросинтеза, основанное на предложенной модели

Основная литература I

Bruno Bienfait and Peter Ertl.

Jsme: a free molecule editor in javascript.

Journal of cheminformatics, 5(1):24, 2013.

Thomas N Kipf and Max Welling.

Semi-supervised classification with graph convolutional networks.

arXiv preprint arXiv:1609.02907, 2016.

Daniel Mark Lowe.

Extraction of chemical structures and reactions from the literature.

PhD thesis, University of Cambridge, 2012.

RDKit.

RDKit: Open-source cheminformatics.

`http://www.rdkit.org`.

[Online; accessed 11-April-2013].

Основная литература II

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling.

Modeling relational data with graph convolutional networks.

In *European Semantic Web Conference*, pages 593–607. Springer, 2018.

Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino.

"found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models.

Chemical science, 9(28):6091–6098, 2018a.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, and Alpha A Lee.

Molecular transformer for chemical reaction prediction and uncertainty estimation.

arXiv preprint arXiv:1811.02633, 2018b.

Основная литература III

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin.

Attention is all you need.

In *Advances in neural information processing systems*, pages 5998–6008, 2017.