Построение признаковых пространств в задаче прогнозирования химических реакций

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.

Графовые нейронные сети в задаче синтеза молекул

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.

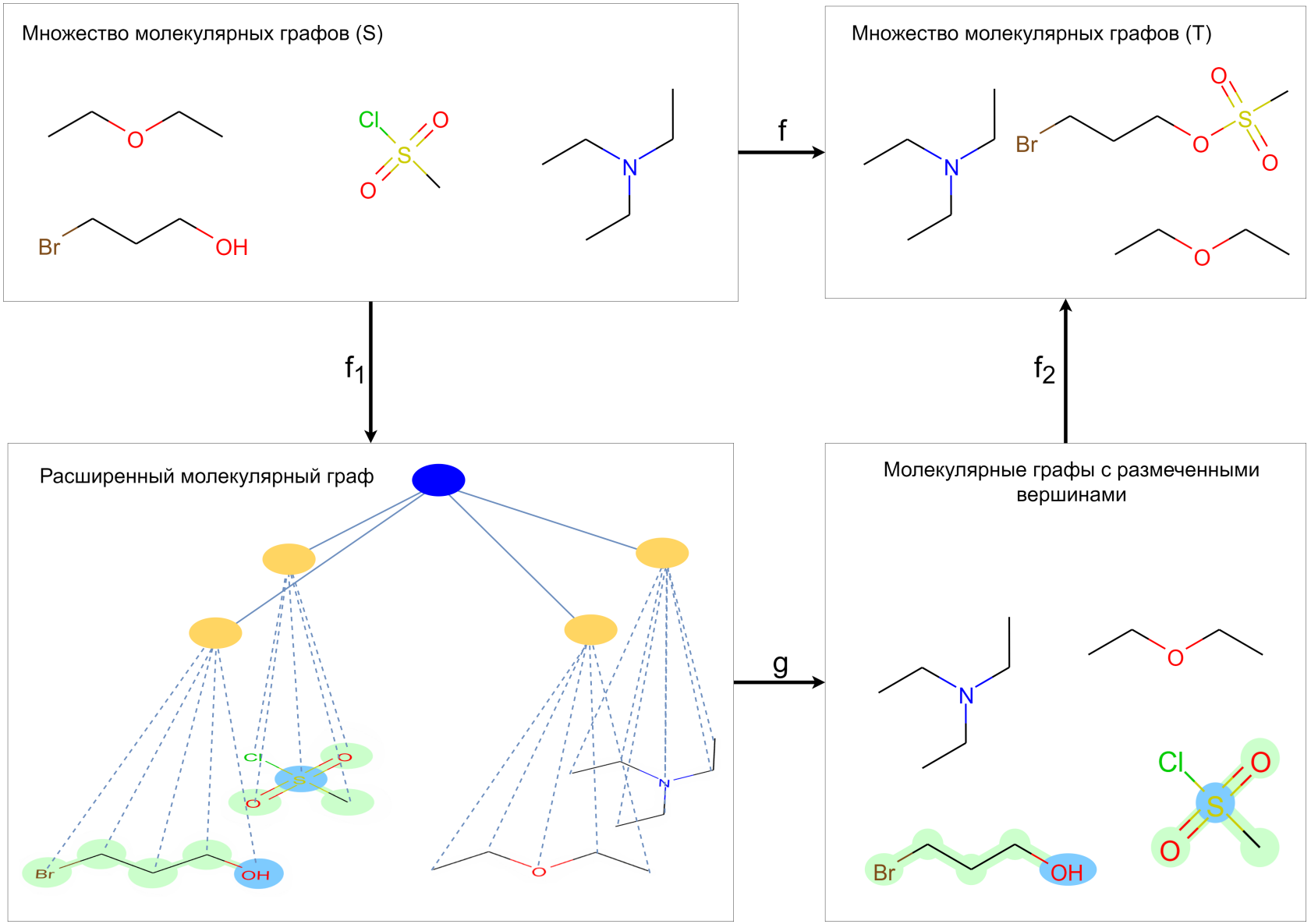
Существующие подходы

1. Butler K. T., Davies D. W., Cartwright H., Isayev O., Walsh A. *Machine learning for molecular and materials science.* Nuture 2018, pp 547-559.
2. Schwaller P., Gaudin T., Lanyi D., Bekas C., Laino T. *"Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-tosequence 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.

Базовые статьи

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.* arXivpreprint arXiv:1709.037412017.

Структура решения



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).

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 $h_i \in \mathbb{H}$, 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 исходных веществ в продукты

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 $\mathcal{F}$ (в работе графовые CNN)

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

где L целевая функция потерь.

Формат исходных данных

База реакций

База реакций из американских патентов США, Lowe, 2012:

1. 1 млн. реакций в формате SMARTS,
2. 1976-2016 год регистрации патента,
3. Разделены катализаторы и реагенты,
4. 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: $\mathrm{c1ccccc[C:1]1[NH2:2]\text{»}c1ccccc[C:1]1[N:2](=O)=O}$

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

RDKit — библиотека для работы с молекулярными графами.

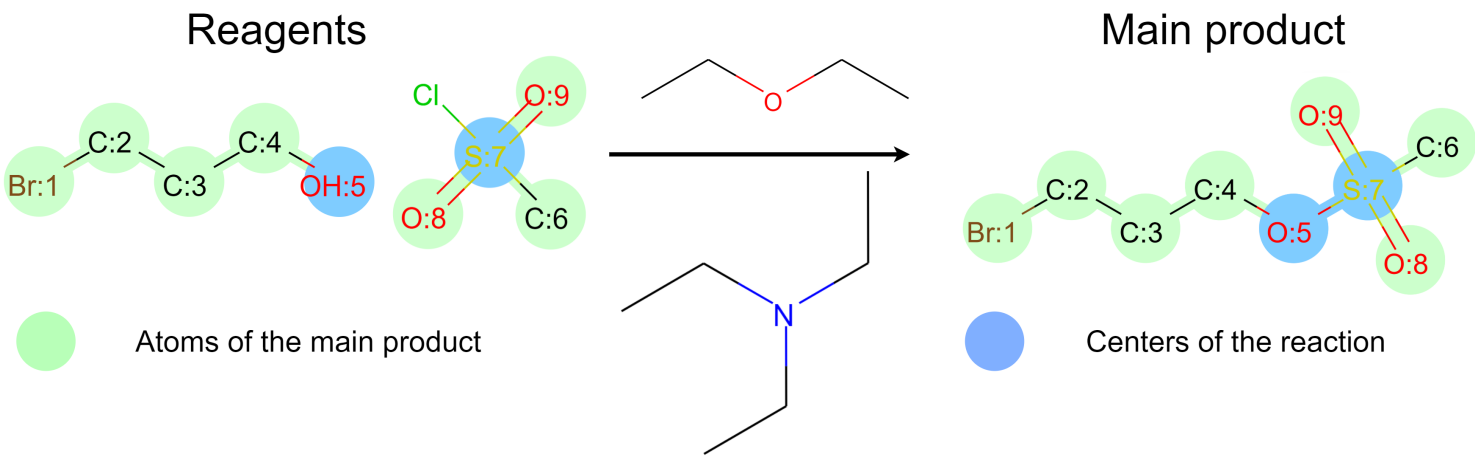
Классификация вершин в графе

Определение атомов основного продукта

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

Определение центров реакции

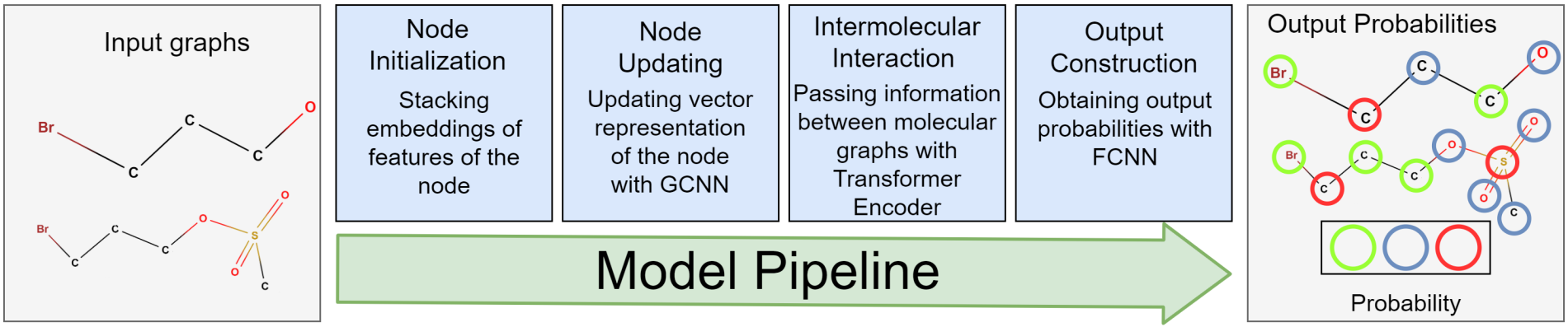
Для множества атомов основного продукта требуется выделить те, конфигурация которых изменилась в течение реакции.



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

Модель классификации вершин в несвязанном графе

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



Relational Graph Convolution Neural Network

Формула обновления состояний вершин

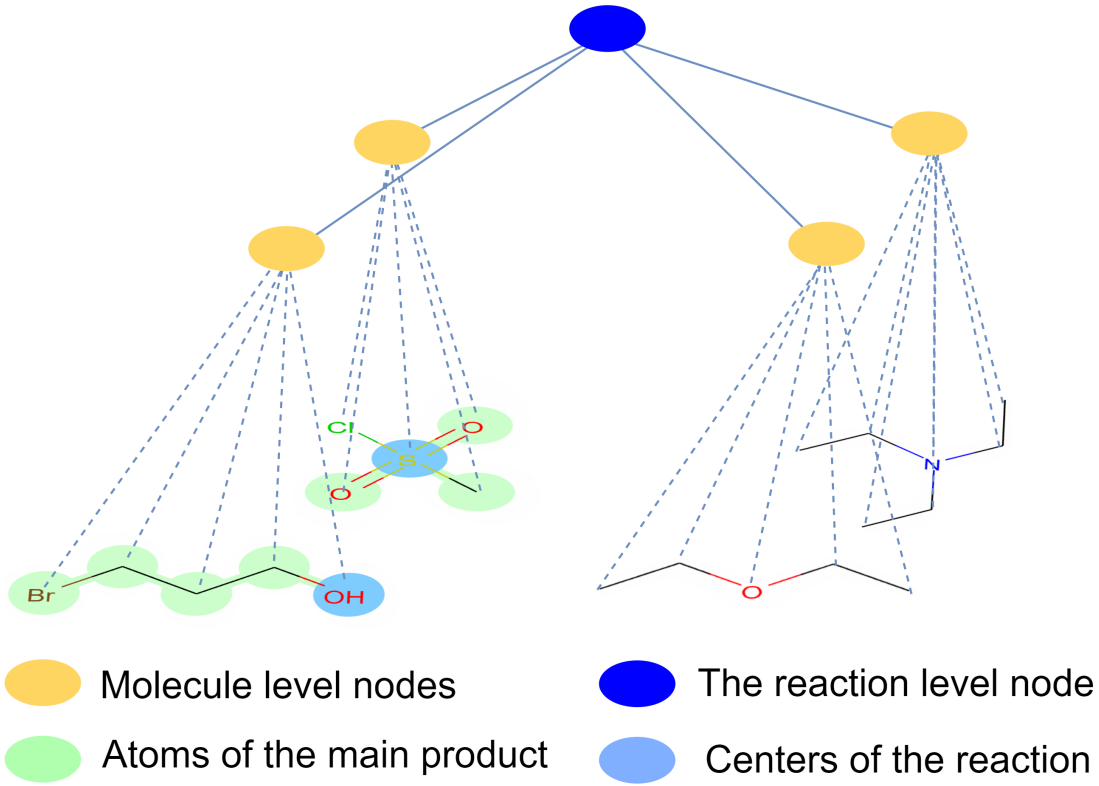
$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right),$$

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

Проблема

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

Расширенный граф химической реакции



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

Расширенный граф химической реакции

Обновление векторных состояний вершин

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

$$\mathbf{h}_{m_k}^{(l+1)} = \text{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}_{rl}^{(l)}\mathbf{h}_r^{(l)} + \sum_{j\in m_k}\frac{1}{|m_k|}\mathbf{W}_{ml}^{(l)}\mathbf{h}_j^{(l)}\right),$$

$$\mathbf{h}_r^{(l+1)} = \text{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_r^{(l)} + \sum_{m_j\in M}\frac{1}{|M|}\mathbf{W}_{rl}^{(l)}\mathbf{h}_{m_j}^{(l)}\right).$$

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

Моделирование взаимодействия атомов разных молекул

Self-Attention

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

Transformer

$$\begin{aligned}
\mathbf{H}^{(l)}_{\text{mha}} &= \text{concat}[\text{head}_1, \text{head}_2, \ldots, \text{head}_h]\mathbf{W}^O, \\
\text{head}_i &= \text{Attention}\left(\mathbf{H}^{(l)}\mathbf{W}^Q_i, \mathbf{H}^{(l)}\mathbf{W}^K_i, \mathbf{H}^{(l)}\mathbf{W}^V_i\right).
\end{aligned}$$

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

Вычислительный эксперимент

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 резльтатов

	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

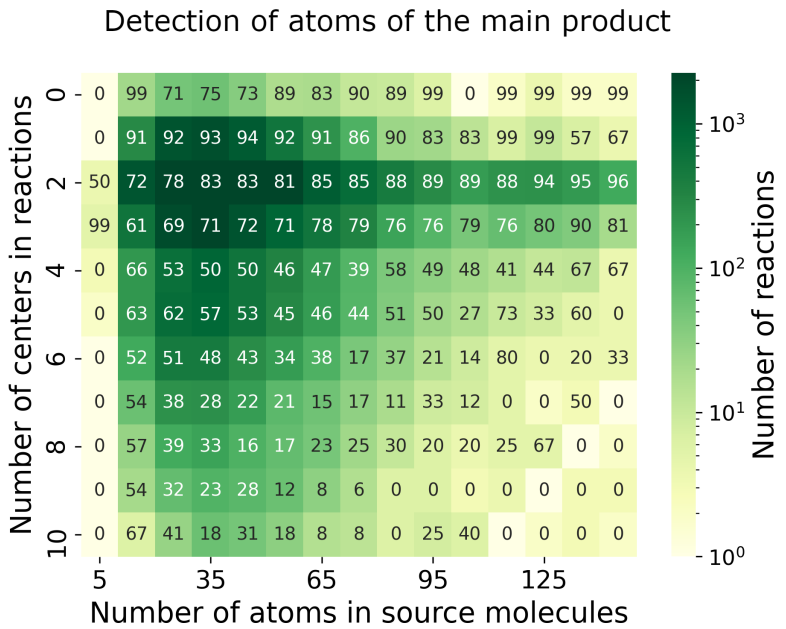
FM среднее значение точности полного совпадения (1, если все метки атомов в реакции предсказаны верно, 0 иначе). F_1 среднее значение F_1-меры между предсказанными и правильными метками атомов в реакции.

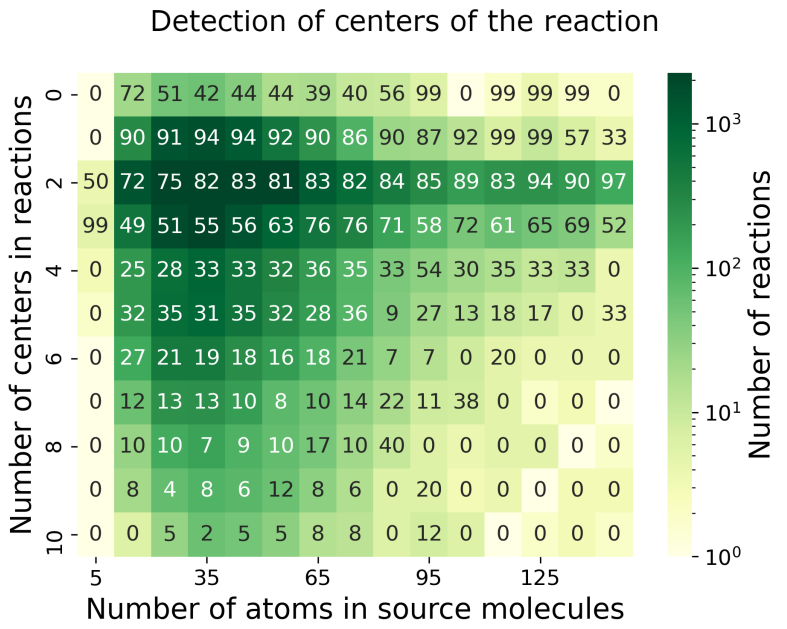
Вывод

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.



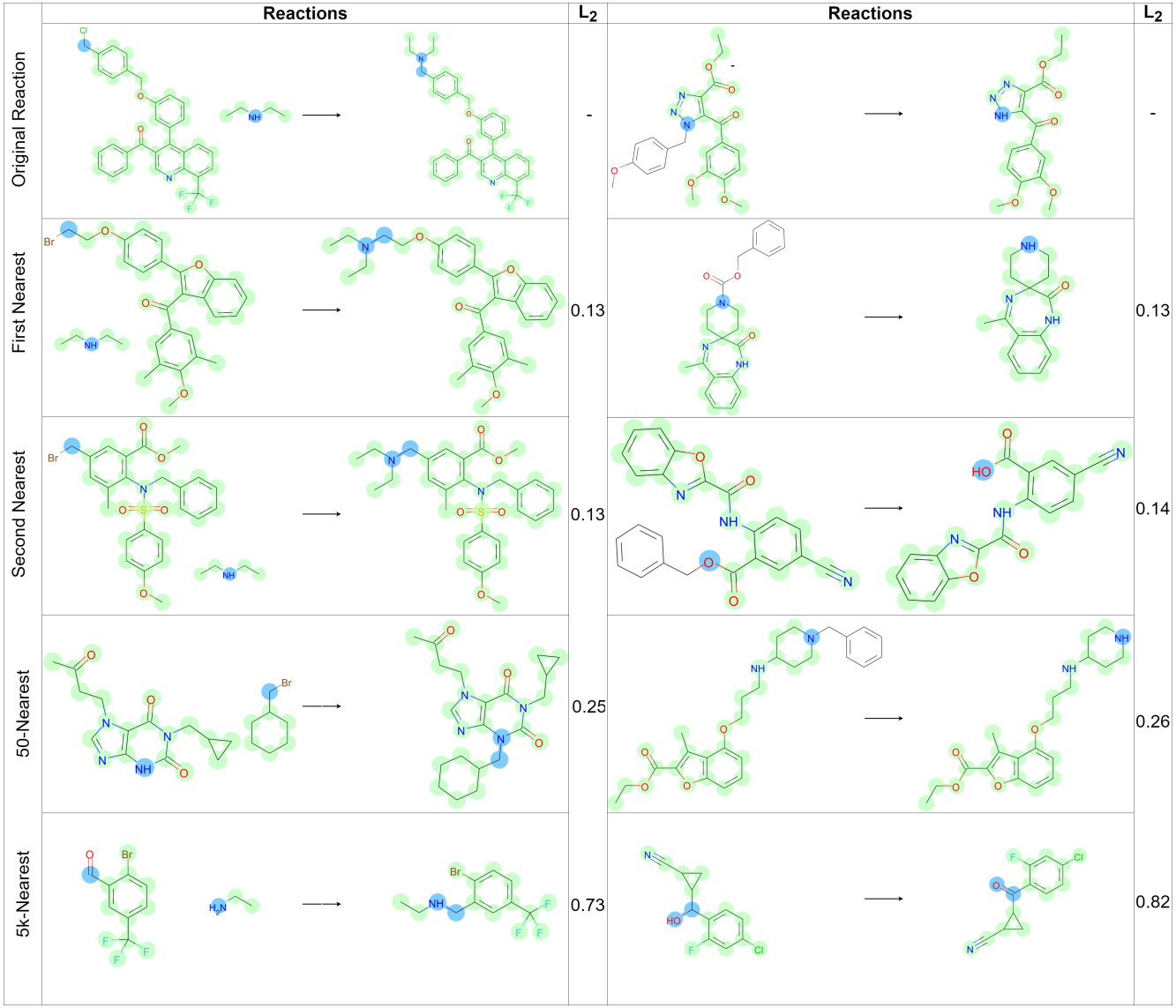


50 The number means full-match accuracy

The color means the number of reactions

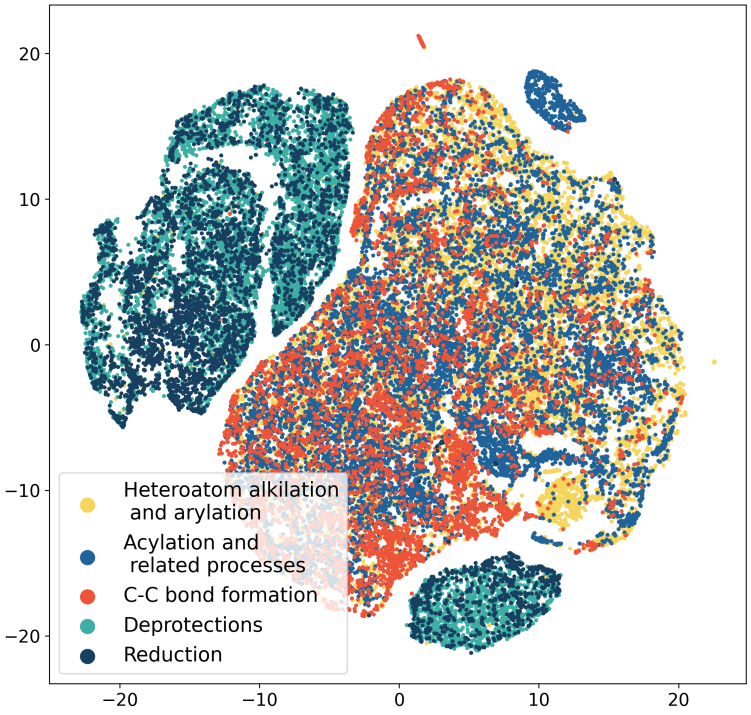
1. С увеличением числа центров качество падает;
2. Зависимость качества от длины исходных молекул выражена слабо.

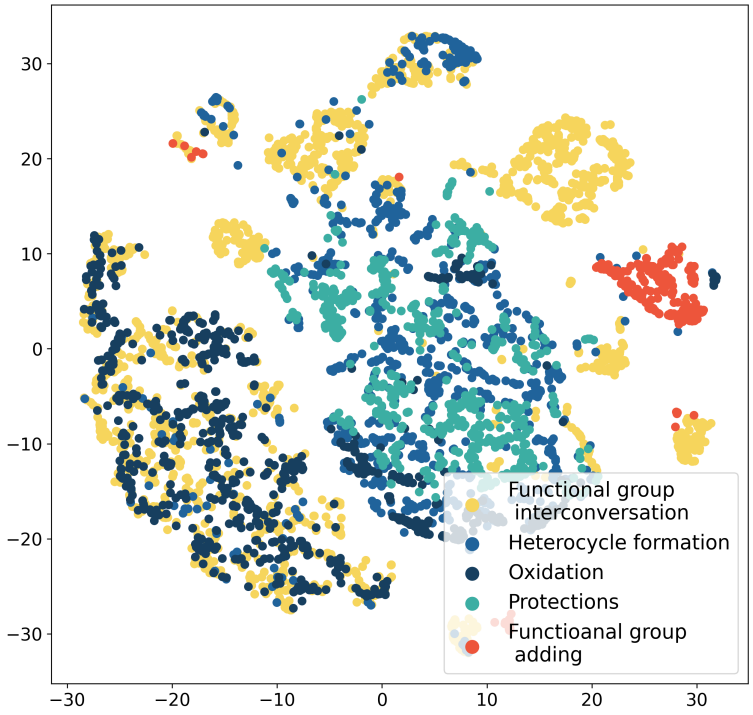
Исследование свойств пространства реакций



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.

Результаты выносимые на защиту

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.

Материалы

1. GitHub репозиторий с задокументированными исходными файлами вычислительных экспериментов.
2. Web-интерфейс предложенной модели, интерактивное представление t-SNE карт реакций.

Опубликованные статьи

Evgeny Egorov, Filipp Nikitin, Vasiliy Alekseev, Alexey Goncharov, and Konstantin Vorontsov.

Topic modelling for extracting behavioral patterns from transactions data.

In *2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI)*, pages 44–444. IEEE.

Filipp Nikitin and Vadim Strijov.

Graph neural network learning for chemical compounds synthesis.

In *Mathematical Methods of Pattern Recognition conference (MMPR-19)*, pages 311–312. Russian Academy of Sciences, 2019.

Filipp Nikitin, Vladimir Dokholyan, Ilia Zharikov, and Vadim Strijov.

U-net based architectures for document text detection and binarization.

In *International Symposium on Visual Computing*, pages 79–88. Springer, 2019.

Ilia Zharikov, Filipp Nikitin, Ilia Vasiliev, and Vladimir Dokholyan.

Ddi-100: Dataset for text detection and recognition.

arXiv preprint arXiv:1912.11658, 2019.