Прогнозирование количественного выхода химических реакций с помощью графовых нейронных сетей.

Гунаев Руслан [1]***, Никитин Филипп***[1]***, Стрижов Вадим***[1]
gunaev.rg@phystech.edu
Московский физико-технический институт

Решается задача определения количественного выхода основного продукта химической реакции по молекулам исходных веществ с помощью графовых нейронных сетей. Данная задача является актуальной задачей вычислительной химии, в частности является подзадачей автоматизации синтеза химических веществ. Предложенная модель использует экспертные знания для решения задачи. Качество решения задачи прямого синтеза измеряется по коэффициенту детерминации между предсказанным и реальным количественным выходом химической реакции.

Ключевые слова: *ретросинтез маленьких молекул, молекулярный граф, выход химической реакции.*

1 Введение

В работе решается задача регрессии для количественного определения выхода основного продукта химической реакции. Исходные вещества и реагенты могут быть представлены в виде нескольких молекулярных графов. В результате взаимодействия этих веществ образуются новые вещества. Представленный набор молекулярных графов исходных веществ и реагентов представляет собой множество неориентированных графов с различными типами ребер и вершин. Тип вершины – номер химического элемента, тип ребра – тип химической связи. Для решения этой задачи авторы используют решения других задач. Например, задачи создания молекулярных структур, таких как молекулярные графы. Эти молекулярные структуры обладают полезными свойствами: они не токсичны и взаимодействуют с другими соединениями. В работе [16] обозрены существующие концепты, нацеленные на использование вычислительных ресурсов для создания молекулярных структур. Предприняты попытки генерировать предположения с помощью рекурентных нейронных сетей [8, 12], которые посимвольно генерируют строковое представление молекулы SMILES [20]. Также для этой задачи применяются вариационные автокодировщики [4] и графовые нейронные сети [5, 6].

Вторая используемая задача заключается в получении предлагаемой молекулы из уже ранее полученных веществ путем реализации цепочки реакций. При этом число реакций в цепочке не должно быть велико. Первые попытки автоматизировать решение задачи синтеза и ретросинтеза веществ были предприняты в 1969 [3]. Сейчас активно развиваются методы глубинного обучения для решения поставленной задачи: предложен вариант использования машинного перевода [1]: продемонстрированы результаты модели перевода с рекурентными слоями [17], а также Трансформер модели [18]. В работе [2] построена графовая сверточная сеть, которая оценивает вероятность образования связи для заданных вершин.

В данной работе решается задача нового типа. В ней требуется по описаниям молекул исходных веществ определить количественный выход химической реакции, как результата химического взаимодействия исходных веществ. В качестве представления молекул используются молекулярные графы. Молекулярный граф – это связный неориентированный граф, находящийся во взаимно-однозначном соответствии со структурной формулой

химического соединения таким образом, что вершинам графа соответствуют атомы молекулы, а рёбрам графа — химические связи между этими атомами.

В работе предложен метод регрессии в молекулярных графах исходных веществ. Каждая реакция – это молекулярный граф, а все входящие в нее молекулы – компоненты. Решение базируется на модели реляционной графовой сверточной нейронной сети(RGCNN) [15]. Предложено несколько модификаций, позволяющих использовать модель для регрессии в несвязанном графе. Модель RGCNN является обобщением модели графовой нейронной сети [9] для графов с различными типами ребер. Такое обобщение является необходимым, так как ребрами в молелярном графе являются химические связи: одинарная, двойная, тройная ароматическая.

2 Данные и их представление

В данной работе используется выборка реакций из патентов США [11]. Данная выборка содержит информацию об 1 млн. химических реакций. Основываясь на данной работе также были получены выборки, являющиеся подмножествами исходного: USPTO_MIT, USPTO_LEF, USPTO_STEREO. Выборки получились в результате фильтрации исходной выборки с различными параметрами в процессе разработки предшествующих алгоритмов. Наибольшая из выборок: USPTO_STEREO содержит 1 млн. реакций, в нем сохранена стереометрическая информация.

Поле	**Описание**	**Пример**
Source	SMILES представления исходных молекул.	`CS(=O)(=O)Cl.OCCCBr>CCN(CC)CC.CCOCC`
Target	SMILES представление основного продукта.	`CS(=O)(=O)OCCCBr`
Canonicalized Reaction	SMILES представление химической реакции.	`CS(=O)(=O)Cl.OCCCBr>CCN(CC)CC.CCOCC>CS(=O)(=O)OCCCBr`
Original Reaction	SMARTS [7] представление химической реакции.	`[Br:1][CH2:2][CH2:3][CH2:4][OH:5].[CH3:6][S:7](Cl)(=[O:9])=[O:8].CCOCC>C(N(CC)CC)C>[CH3:6][S:7]([O:5][CH2:4][CH2:3][CH2:2][Br:1])(=[O:9])=[O:8]`
Patent Number	Уникальный номер патента	US03930836
Paragraph Number	Номер параграфа	2
Year	Год публикации	1976
Mined Yield	Количественный выход реакции	75%

Таблица 1 USPTO_STEREO выборка химических реакций.

Для получения графового представления молекул в реакции использовался программный пакет RDKit [14]. С его помощью из исходных SMARTS представлений были получены: исходные типы атомов молекулы, признаки атомов, матрицы смежности с указанием типа химической связи.

3 Постановка задачи

Дана выборка – $X = \{g_i, y_i\}_{i=1}^N$, где $g_i \in G$ – множество исходных молекулярных графов, $y_i \in [0; 1]$ – выход реакции. Рассмотрим модель

$$f(g, \mathbf{W}) : G \times \Omega \to \mathbb{R}, \tag{1}$$

где $\mathbf{W} \in \Omega$ – пространство параметров модели.

Требуется найти такие параметры модели $\mathbf{W}^*$, что

$$\mathbf{W}^* = \arg\min_{\mathbf{W}\in\Omega} \frac{1}{N} \sum_{i=1}^{N} \|y_i - f(g_i, \mathbf{W})\|_2^2 \tag{2}$$

3.1 Архитектура модели

Пайплайн модели состоит из следующих этапов: инициализация векторного представления вершин графа, обновление представлений, основываясь на структуре графа, сбор информации с различных компонент графа, построение выхода, оценка функции ошибки.

3.2 Начальное представление атома

Каждый атом представлен несколькими категориальными признаками, такими как валентность, тип атома и др. Признаки обозначим $k_n, n \in \{1, \dots M\}$.

Для каждого из представленных категориальных признаков строится векторное представление (embedding). В формуле (3) используются обозначения: i — номер вершины в графе, m_i — тип вершины, k — номер признака, W^k — матрица весов для k-го категориального признака, W_i^k — i-ый столбец матрицы.

$$\mathbf{h}_{ik}^{(0)} = \mathbf{W}_i^k. \tag{3}$$

Преставление атома есть конкатенация эмбедингов $\mathbf{h}_{ik}^{(0)}$ по всем признакам (4).

$$\mathbf{h}_i^{(0)} = \text{concat}[\mathbf{h}_{i0}^{(0)}, \mathbf{h}_{i1}^{(0)}, \mathbf{h}_{i2}^{(0)}, \dots, \mathbf{h}_{iM}^{(0)}]. \tag{4}$$

3.3 Обновление скрытых состояний атома

В графовой сверточной нейронной сети обновление скрытых состояний происходит в соответствии с формулой (5). В этой формуле: N_i — множество индексов атомов смежных с i-м, c_i — нормализационный коэффициент.

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{j \in N_i} \frac{1}{c_i}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right). \tag{5}$$

Недостатком данной модели является предположение о том, что все связи между вершинами идентичны(каждое смежное ребро умножается на одну матрицу с одинаковым коэффициентом). Для предсказаний высокой точности необходимо, чтобы модель выучивала типы связей в процессе обучения. Эту проблему решает RGCNN, обновление скрытых состояний в которой происходит в соответствии с формулой (6). В данной модели смежные с заданной вершиной разным типом связи атомы учитываются с разными весами.

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \sum_{r \in R}\sum_{j \in N_i} \frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right). \tag{6}$$

3.4 Введение скрытых представлений молекулы и реакции

Предлагается ввести скрытые векторные представления молекул исходных веществ $\mathbf{h}_{m_i}^{(l)}$ и векторное представление химической реакции $\mathbf{h}_r^{(l)}$. Векторное представление атома в молекуле связано с соответственным векторным представлением молекулы, а векторное представление молекулы связано с векторным представлением реакции. Создание векторного представления молекулы при использовании графовых сверточных сетей было предложено в работе [10]. В работе предлагается обобщение данного метода на случай, когда входные данные представляют собой несколько молекулярных графов.

Модель RGCNN позволякт использовать различные веса обновления для различных типов связи. Придем к следующим формулам обновлений векторных представлений (7), (8), (9).

$$\mathbf{h}_i^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_i^{(l)} + \mathbf{W}_{\mathrm{ml}}^{(l)}\mathbf{h}_{m_k}^{(l)} + \sum_{r\in R}\sum_{j\in N_i}\frac{1}{c_{i,r}}\mathbf{W}_r^{(l)}\mathbf{h}_j^{(l)}\right), \tag{7}$$

$$\mathbf{h}_{m_k}^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}_{\mathrm{rl}}^{(l)}\mathbf{h}_r^{(l)} + \sum_{j\in m_k}\frac{1}{|m_k|}\mathbf{W}_{\mathrm{ml}}^{(l)}\mathbf{h}_j^{(l)}\right), \tag{8}$$

$$\mathbf{h}_r^{(l+1)} = \mathrm{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_r^{(l)} + \sum_{m_j\in M}\frac{1}{|M|}\mathbf{W}_{\mathrm{rl}}^{(l)}\mathbf{h}_{m_j}^{(l)}\right), \tag{9}$$

где $\mathbf{W}_{\mathrm{ml}}$– матрица весов, соответствующая связи между атомом и молекулой, $\mathbf{W}_{\mathrm{rl}}$ – матрица весов для связи между молекулой и реакцией, $|m_k|$ – количество атомов в молекуле, $|M|$ – количество молекул в реакции.

Введение таким образом векторного преставления реакции и молекул означает переход от исходного множества графов исходных веществ к единому графу с дополнительными вершинами, отвечающими за данные преставления(см. Рис. 1). Это также означает, что связи на уровне молекулы и реакции имеют разный тип.

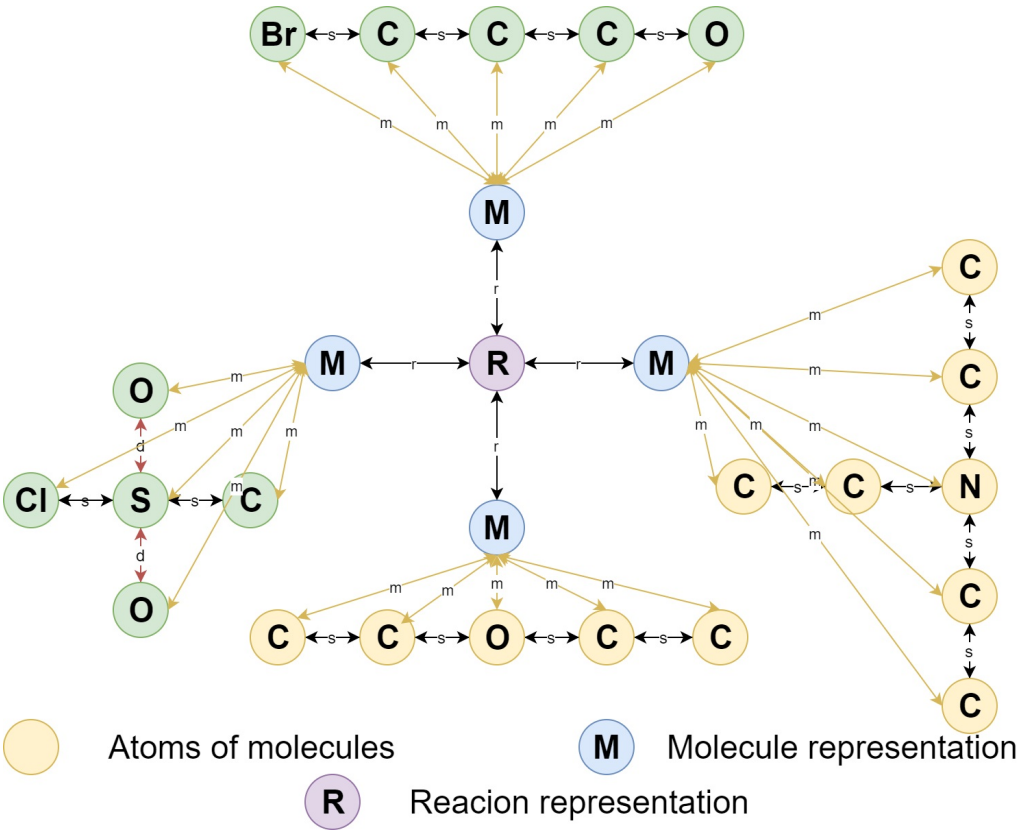


Рис. 1 Расширенный граф с введенными вершинами — представениями молекул и реакции

3.5 Агрегация графа.

После прохождения RGCNN каждой реакции соответвует определенное число атомов, пусть n, тогда сделаем агрегацию графа.

$$\mathbf{h}_g = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_i^{(m)}, \tag{10}$$

где m – количество слоев в RGCNN.

3.6 Построение выхода сети

Полученные векторные представления графов подаются на вход полносвязанной нейронной сети, решающей задачу регрессии (12)

$$\mathbf{h}_g^{(l+1)} = \mathrm{ReLU}(\mathrm{linear}(\mathbf{h}_g^{(l)})) \tag{11}$$

$$\hat{y} = \mathrm{linear}(\mathbf{h}_g^{(t)}), \tag{12}$$

где t – количество слоев в полносвязанной нейронной сети.

3.7 Функция ошибки

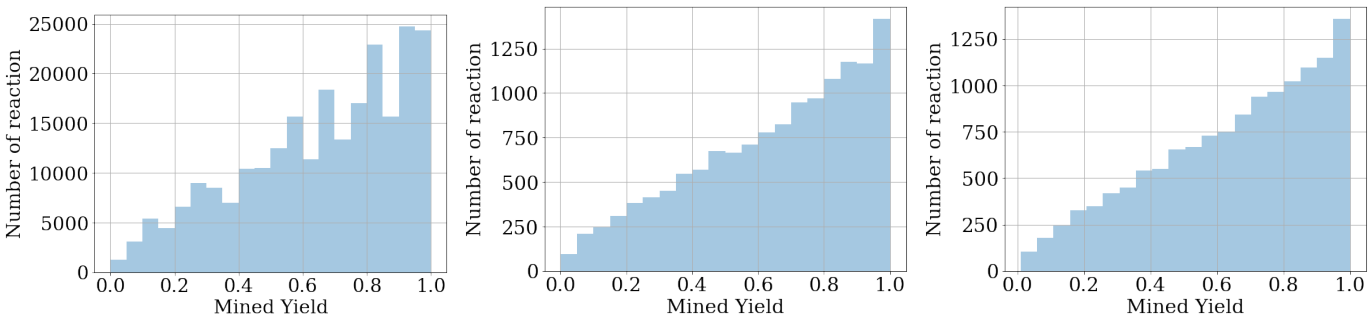
В качестве функции ошибки была использована MSELoss, значение которой вычисляется следующим образом (13).

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\|y_i - \hat{y}_i\|_2^2 \tag{13}$$

4 Вычислительные эксперименты.

Авторами редлагается несколько архитектур. Авторы приследуют две основные цели. Первая – проверить, повышается ли качество модели при добавлении дополнительной информации о структуре графа. Вторая – сравнить качество с с константной моделью.

Для начала проведения экспериментов необходимо найти в используемых выборках реакции, в которых известно поле **Mined Yield**. В результате этого каждая из выборок (train, test, valid) сокращается на 80%. Поставленная задача является трудной. Во-первых, единственная доступная выборка содержит примеры всевозможных типов реакций. Во-вторых, после преобразований выборки становятся несбалансированными и неоднородными (см. Рис. 2). В-третьих, задача решается впервые, поэтому для адекватной оценки модели в качестве простейшей выбрана константная **CONST** модель.



(*а*) Train (*б*) Test (*в*) Valid

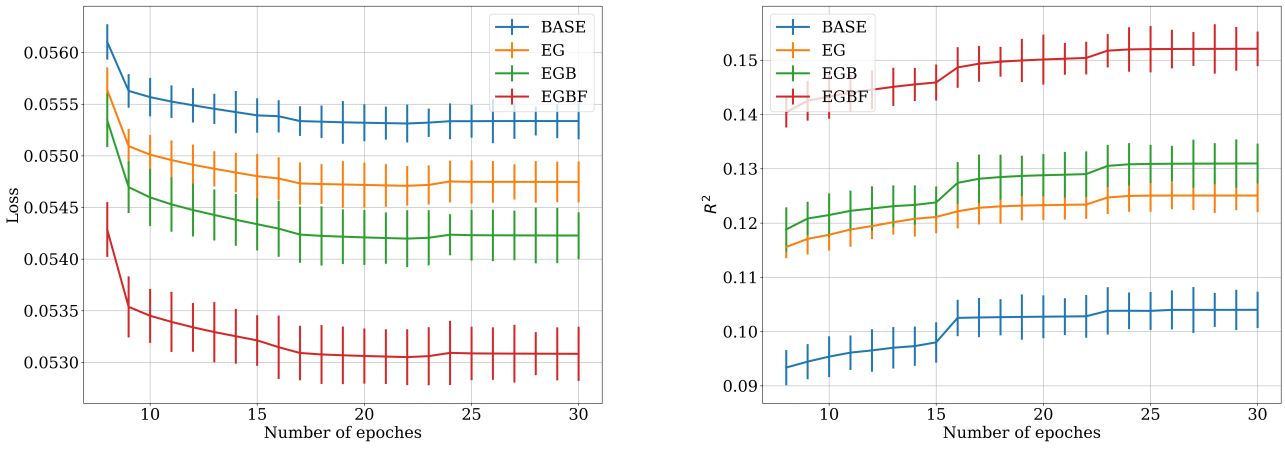
Рис. 2 Распределение выходов реакций в каждой из выборок.

Таблица 2 Результаты экспериментов. R^2 – коэффициент детерминации. MAE – среднее абсолютное отклонение между реальными выходами реакций и предсказанными на тестовой выборке. Коэффициент детерминации для константной модели не определен.

Модель	**Mined Yield**	
	R^2	MAE
CONST	0	0.211 ± 0.004
BASE	0.104 ± 0.002	0.198 ± 0.003
EG	0.125 ± 0.003	0.194 ± 0.002
EGB	0.131 ± 0.006	0.186 ± 0.002
EGBF	0.152 ± 0.005	0.174 ± 0.003

4.1 Эксперименты

Базовая модель **BASE** состоит из двух частей: **RGCNN** и **FCNN** и принимает несвязанный граф исходных молекул. Предполагается, что известны только типы вершин. Данная модель не использует никакой специфичной информации (типы связей, свойства атомов) для конкретной задачи и не способна к передаче данных между компонентами(молекулами) в несвязанном графе. Модель демонстрирует самое низкое качество среди всех поставленных экспериментов(см. табл. 2), потому что конечное представление атома зависит только от представлений атомов в этой же молекуле. Однако основным механизмом химических реакций является именно межмолекулярное взаимодействие.



(*а*) Loss (*б*) R^2

Рис. 3 Результаты базового (**BASE**) эксперимента.Loss – функция ошибки в процессе обучения, R^2 – коэффициент детерминизации в процессе обучения между реальными значениями выхода тестовой выборки и предсказанными.

Использование расширенного молекулярного графа **EG** в качестве входа **RGCNN** повышает качество модели. Повышение качества показывает, что дополнительное представление молекул и всей реакции моделирует межмолекулярное взаимодействие в химической реакции и обменивает информацию по исходным молекулярным графам.

Следующие модификации архитектуры заключаются в предоставлении дополнительной информации об атомах и связях. Эмбединги вершин содержат информацию о различных свойствах атома. Кроме того, реляционная структура сверточных слоев графа имитирует различные типы химических связей.

Модель **EGB** работает с различными типами химических связей: одинарными, двойными,тройными, ароматическими. Это приводит к повышению качества. Наибольшее влияние на конечный результат оказывает использование различных расчетных свойств атомов при инициализации эмбедингов вершин в расширенном молекулярном графе. Свойства: степень, явная валентность, гибридизация, неявная валентность, ароматичность, неявность, число явных водородов, число неявных водородов, кольцо, число радикальных электронов, формальный заряд. Добавление этих свойств в модель **EGBF** повышает качество модели.

Все эксперименты были проведены при помощи библиотек PyTorch [13] и DGL [19], запущены на Nvidia 1080 Ti. 30 эпох обучения на лучшей архитектуре занимало примерно 6 часов.

5 Выводы и планируемые исследования

В данной работе построена регрессионная модель на множестве молекулярных графов. Проделанные эксперименты показали, что информация о структуре молекулярного графа повышает качество модели. Каждая из представленных архиктектур показала более высокое качество, чем константная модель. Дальнейшее улучшение модели возможно за счет обучения на выборке, в которой, во-первых, есть разделение на типы реакции, во-вторых, равномерное распределение выходов реакций. Также для увеличения качества регрессии предлагается классифицировать реакции по их типам. .

Литература

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, WH Green, R Barzilay, KF Jensen, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. 2018.

[3] EJ Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses. *Science*, 166(3902):178–192, 1969.

[4] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[5] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[6] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[7] Ward Edwards and F Hutton Barron. Smarts and smarter: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes*, 60(3):306–325, 1994.

[8] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[10] Junying Li, Deng Cai, and Xiaofei He. Learning graph-level representation for drug discovery. *arXiv preprint arXiv:1709.03741*, 2017.

[11] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[12] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

[13] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[14] RDKit: Open-source cheminformatics. `http://www.rdkit.org`. [Online; accessed 11-April-2013].

[15] 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.

[16] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649, 2005.

[17] 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, 2018.

[18] 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*, 2018.

[19] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. Deep graph library: Towards efficient and scalable deep learning on graphs. *arXiv preprint arXiv:1909.01315*, 2019.

[20] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Поступила в редакцию