Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Никитин Филипп Александрович

Построение признаковых пространств в задачах прогнозирования химических реакций

03.04.01 — Прикладные математика и физика

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2020

Содержание

Аннотация

Методы машинного обучения были применены к ряду задач построения компьютерных ассистентов предсказания синтеза(CAST). Задача предсказания продуктов химических реакций была сформулирована в терминах классификации вершин несвязанного графа исходных веществ химической реакции. Предложено обобщение графовой сверточной сети для работы с несвязанными графами. Продемонстрировано, что предложенный подход успешно предсказывает продукт реакции и отображение атомов исходных веществ в атомы основного продукта химической реакции. Эволюционное семейство вычислительных экспериментов показывает влияние всех предложенных модификаций на итоговое качество рассматриваемой модели на датсете реакций, извлеченных из патентов США(USPTO). Дополнительный анализ модели демонстрирует её интерпретируемость. Введенные латентные векторные представления реакций неявно образуют скоррелированные с типом химической реакции кластеры, метрически близким векторам в данном латентном пространстве соответствуют реакции с похожим механизмом протекания. Программная реализация предложенных алгоритмов оформлена в виде библиотеки с открытым исходным кодом.

Ключевые слова: *графовая сверточная нейронная сеть, механизм внимания, химическая реакция, несвязанный граф, классификация вершин графа.*

Введение

Актуальность темы. Магистерская работа посвящена задаче прогнозирования продуктов химических реакций. В общем виде задача может быть рассмотрена как поиск отображения множества графов во множество графов. Где исходными графами являются молекулярные графы исходных веществ в химической реакции, а результат отображения — множество молекулярных графов продуктов. По имеющейся базе химических реакций большого объема требуется построить модель, которая по молекулярным графам исходных веществ предсказывает молекулярные графы продуктов. Важнейшим требованием к модели является её обобщающая способность — возможность модели работать с данными не представленными в процессе её построения.

Решение задачи актуально в области автоматизации синтеза химических веществ и открытии потенциальных лекарственных препаратов [1–4]. В описанных работах обоснован потенциал методов машинного обучения для решения задачи построения компьютерных ассистентов предсказания синтеза(CAST) и актуальность совершенствования методов в современной вычислительной химии.

В последние несколько лет было представлено ряд работ, направленных на решение задачи CAST. Часть из них оперирует со строковыми представлениями графов [5–7]. Несмотря на то, что представленные методы демонстрируют наилучшее качество, имеется ряд серьезных проблем, ограничивающих их потенциал и область применимости. Методы используют подходы, предложенные ранее для работы с последовательностями символов и машинного перевода. Данные методы не являются интерпретируемыми, вход и выход модели представим только в виде последовательности категориальных признаков. Это затрудняет использование локальных признаков атомов и химических связей в молекулярном графе. Графовое представление молекулы имеет ряд преимуществ: в нем явно задаются вершины и ребра, признаки и свойства которых изучаются экспертами, могут быть вычислены и использованы в модели. Была предложена модель, использующая графовую сверточную нейронную сеть и вычисляющая попарные вероятности образования химической связи между вершинами [8]. Данный метод имеет интерпретируемую архитектуру, однако не является универсальным методом работы с несвязанными графами. Другое направление исследований сосредоточено на построение систем формальных правил, согласно которым определяется преобразование исходных молекулярных графов в продукт химической реакции [9–13]. Несмотря на полную интерпретируемость данных моделей, они не обладают необходимой обобщающей способностью и не могут быть применены к химическим реакциям, механизм которых не описан. С ростом числа открытых химических соединений, возникли трудности с описанием формальных правил, так как их количество сильно росло и не обладало полнотой.

Таким образом среди существующих методов предсказания продуктов химических реакций нет интерпретируемых, универсальных методов с достаточной обобщающей способностью. Отсутствуют подходы машинного обучения, позволяющие эффективно оперировать с данными, представимыми в виде несвязанных графов.

Цель работы. Целью работы является разработка метода предсказания основного продукта химических реакций по молекулярным графам исходных веществ

- применим к данным в виде несвязанного молекулярного графа;

- допускает использования экспертных знаний о локальной структуре молекулярного графа;

- демонстрирует адекватные результаты на выборке реакций большого объема;

теоретическое обоснование свойств разработанного метода, а также построение грамотного и объективного вычислительного эксперимента.

Методы исследования. Для достижения целей используется аппарат теории оптимизации, машинного обучения, глубинного обучения, линейной алгебры. Для программной реализации использовался язык программирования Python 3.7, вычислительные эксперименты проводились на кластере.

Основные положения, 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.

Научная новизна. Сформулирована задача предсказания продуктов химических реакций в терминах бинарной классификации вершин несвязанного графа. Предложено два алгоритма обобщения графовой сверточной нейронной сети для несвязанных графов.

Теоретическая значимость. Данное исследование является базовым в обобщение методов машинного обучения для графовых представлений данный на случай несвязанных графов.

Практическая значимость. Результаты экспериментов показывают сравнимое с существующими моделями качество, интерпретируемость предложенной модели является существенным преимуществом для практического применения в системах автоматического синтеза химических элементов. Вычислительный эксперимент проводился на большой выборке химических реакций из патентных заявок.

Степень достоверности и апробация работы. Достоверность результатов подтверждает изложенная математическая часть, вычислительный эксперимент, анализ модели. Работа подана к публикации в рецензируемый научный журнал. Промежуточные результаты работы докладывались и обсуждались на следующих научных конференциях

- 19-я Всероссийская конференция с международным участием «Математические методы распознавания образов» [14], 2019;
- Ежегодный Саммит молодых ученых и инженеров «Большие вызовы для общества, государства и науки», 2019.

Literature review

Drug discovery and development pipelines are time-consuming, resource-intensive, and sophisticated. The development of a new drug takes several years and costs billions of dollars [15, 16]. Therefore, automation of the process is an actual and important problem [17]. Early drug development consists of two processes: drug discovery and retrosynthetic analysis [16]. In the first process, molecules, which possess suitable characteristics to make acceptable drugs, are identified. The second process is target-oriented syntheses of these molecules. The synthesis can be planned effectively with retrosynthetic analysis [18]. The goal of the analysis is to construct a synthesis path from buyable compounds to the small molecules. The path should be short in order to obtain a valuable amount of the target substance.

Drug discovery can be viewed as a challenging multi-dimensional problem in which various characteristics of compounds need to be optimized together to provide drug candidates. The idea for a target can come from a variety of sources, including academic and clinical research. Recently, advances in computer science have changed the drug discovery process [1, 19]. There are several works aimed at the prediction of new chemical compounds with given characteristics, including efficacy, pharmacokinetics, and safety. Concepts of using recent advances in computer science are proposed to generate new targets automatically [20]. Drug discovery is a computationally hard problem because the space of available molecules is huge [21].

Recent trends in data science demonstrate the potential of machine learning and deep learning technique to solve a wide range of different problems in a variety of fields such as natural language processing, computer vision, and signal processing [22–24]. Moreover, the influence of the methods in computational chemistry and biology was displayed in the paper [1–4]. Models for a generation of drug candidates were built with a recurrent neural network that generates SMILES [25] representation of target compound character by character [26,27]. Variational autoencoders and graph neural networks was successfully applied to the problem [28–30].

The synthesis research carried out in universities and applied in laboratories has changed the drug development industry [1,31]. For example, the reactions are asymmetric synthesis and metal-catalyzed cross-coupling reactions [32,33]. Today outcomes prediction is particularly routine for medical chemists. Organic chemists recognized the potential of computational methods in practice and developed the first rule-based method (OCSS) 50 years ago [9]. There were other attempts to develop methods to automate retrosynthetic analysis: CAMEO [10], EROS [11], IGOR [12], SOPHIA [13]. Medical chemists use a huge set of unstructured rules to predict products in the reaction (see Fig. 1). Computer-aided retrosynthesis would be a valuable tool, but at present, it is slow and provides results of unsatisfactory quality [5].

Modern approaches to the problem rely on deep learning methods [5–8]. Architecture of Neural Machine Translation is adapted to the forward synthesis problem. The architecture initially solves natural language translation problem [34]. The methods use SMILES representation of the reagents, reactants, and products. It translates the source string to the product string character by character. Recurrent and Transformer architectures for NMT demonstrate the excellent performance of the task. The last architecture is now the state of the art solution. However, fully data-driven approaches have several weaknesses. The methods do not use any expert knowledge about graph

structure, SMILES language construction. A graph convolution neural network was proposed to evaluate the probability of a bond between two nodes. To sum up, deep learning methods have the potential to solve the problem acceptable at industry.

The sequence-to-sequence models can be learned with SMILES of reagents and products [35]. However, more sophisticated methods require the atom mapping [36]. The atom mapping of a chemical reaction is a bijection of the reactant atoms to the product atoms that specifies the terminus of each reactant atom. Finding the atom mapping in reactions is essential in classifying reactions in large databases, facilitating substructure searches, identifying metabolic pathways [37–39]. One of the problems with existed databases is the inconsistent quality of the data entered into these databases over the years. Most of the reactions are not balanced and not atom-mapped. It alone creates significant problems for automated machine understanding of chemical reactions and reactivity [40].

Problem statement

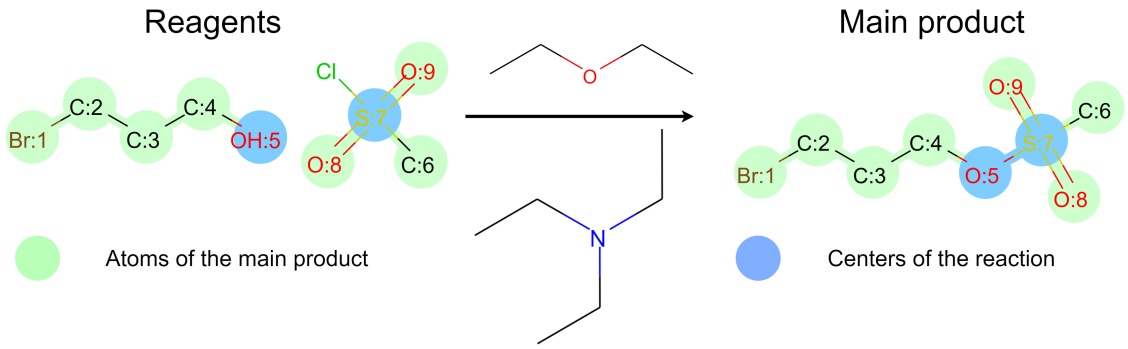


Рис. 1: A chemical reaction maps reagents into products. On the molecules of reagents, two types of atoms are labeled: atoms of the main product and centers of the chemical reaction. Centers are the atoms that change their characteristics.

The authors propose a method that is capable of predicting outcomes and finding atom mapping at the same time. Two specific tasks are solved in parallel (see Fig. 1). Atoms of the main product and centers of the reaction are found. Centers of the reaction are atoms of the main product, which change the configuration in the reaction. The configuration of an atom is a superposition of characteristics of the atom and adjacent bonds. In terms of graph theory, both tasks are node-classification in a disconnected graph of source molecules. The novel neural network MolsNet solves the node-classification tasks. Atoms of the main product and centers of the reaction determine the outcome in the majority of reactions because they have less than three centers.

The method structure (see Fig. 2) consists of several blocks. Firstly, each atom is mapped to a real vector according to its characteristics in the molecule. The model is capable of using any known numerical characteristics of atoms. Secondly, the vectors are updated with Relational Graph Convolution Neural Network (RGCNN) [17]. The RGCNN generalizes Graph Convolution Neural Network [41] for graphs with different edge types that correspond to chemical bonds. The authors offer to use extended molecular graphs with molecule's and reaction's level nodes to enable passing information across different molecules. Then, the Transformer encoder processes the vectors. The block simulates intermolecular interaction, which is a mechanism of chemical reactions. Finally, the Fully-connected neural network (FCNN) gives probabilities for each atom in the node classification problems.

Compared with other recent studies, MolsNet has several novel aspects in terms of architecture of neural network. MolsNet generalizes the graph convolution neural network for the disconnected graph of molecules. The natural structure of the MolsNet is suitable to add information about molecules and atoms: characteristics of atoms, types of chemical bonds.

Experiments are conducted on the dataset of reactions which was extracted from the US patents (see Tab. 1) [42]. The results demonstrate excellent performance. The proposed methods of generalization RGCNN architecture work with disconnected graphs. Additional expert knowledge about the structure of the molecular graph results in increase of model quality. After that, a comprehensive analysis of the best model illustrates that the model learns chemical insights from the given reactions.

Dataset

Field	Description	Example
Source	SMILES of source molecules	CS(=O)(=O)Cl.OCCCBr>CCN(CC)CC.CCOCC
Target	SMILES of the main product	CS(=O)(=O)OCCCBr
Canonicalized Reaction	SMILES of the chemical reaction	CS(=O)(=O)Cl.OCCCBr>CCN(CC)CC.CCOCC> CS(=O)(=O)OCCCBr
Original Reaction	SMARTS of the chemical reaction	[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	Unique number of the patent	US03930836
Paragraph Number	Paragraph number in the patent	2
Year	Year of publication	1976

Таблица 1: The USPTO_STEREO dataset of chemical reactions. The dataset consists of one million chemical reactions extracted from the US patents, which was registered between 1976 and 2015.

Most of the publically available datasets are based on a set of reactions that were extracted from United States patents published between 1976 and September 2016 with text-mining [42]. The original patent information describes a complex chemical synthesis process consisting of multiple steps. The information summarised to a SMARTS [43] string (see Tab. 1), which includes three groups of molecules: the reactants, the reagents, and the products. Any other information about the synthesis process such as a physical condition was removed. The original dataset has noise and duplicate examples. In the previous studies, [6, 8] quality of methods is evaluated on subsets. Reactions without duplicates and with a single product make up the USPTO_STEREO dataset, which contains one million reactions. The USPTO_MIT is obtained with more sophisticated filtering. It consists of 300k reaction. The USPTO_50k contains 50 thousands of reactions which has one of ten classes.

The SMARTS representation of a reaction is converted to a molecular graph with open-source library RDKit [44]. The library is used to calculate atom features: degree, explicit valence, hybridization, implicit valence, aromaticity, implicitness, number of explicit hydrogenous, number implicit hydrogenous, is a ring, number of radical electrons, formal charge.

Method

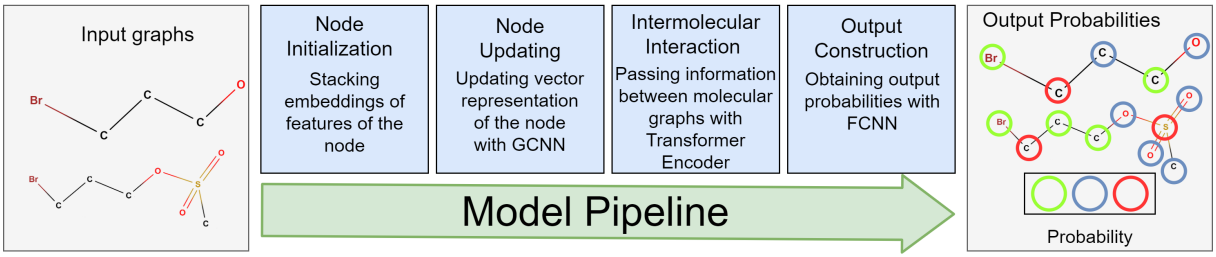


Рис. 2: The MolsNet architecture. Each step in this pipeline naturally corresponds to the structure of given molecular graphs. It uses different local features of atoms in molecules to construct an initial representation of nodes. The final atom representation is given according to the adjacent node and edges, and other molecular graphs in the reaction. Each atom and molecules impact the final probabilities with specific weights.

The core of the proposed method is Graph Neural Network [45]. Therefore, the pipeline consists of the next stages (see Fig. 2): initialization of vector representations of nodes in graphs, updating the representation according to structure of the graphs, aggregation information from different components of the disconnected graph, construction of output, and evaluation loss function.

Initialization of vector representation of nodes. Description of the atom consists of several categorical features such as type of atom, valency, formal charge. For each feature, vector embedding is constructed,

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

Where f_i^k is a value of feature k of atom i, $\mathbf{W}^k$ is an embedding matrix for categorical feature k.

The final vector representation is a concatenation of embeddings of all features,

$$\mathbf{h}_i^{(0)} = \text{concat}[\mathbf{h}_{i0}^{(0)}, \mathbf{h}_{i1}^{(0)}, \mathbf{h}_{i2}^{(0)}, \ldots, \mathbf{h}_{iK}^{(0)}].$$

Updating of vector representation. In graph convolutional neural network, vector representation of nodes is updated according to equation,

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

Where ReLU is a rectified linear unit, N_i is a set of atoms which is adjacent with i, c_i is a normalising factor.

One disadvantage of the model is the assumption that edges in the graph are the same. As we discussed, the type of chemical bond is an important feature in the investigated problem. In Relational Graph Convolution Neural network (RGCNN [17]), a

vector representation of node is updated according to the equation,

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

Where for each type of edge, there is a unique weight matrix.

Passing information between graph components. The output of the model should depend on all atom's representation in source molecules. Therefore, the updating mechanism of RGCNN should be improved to work with disconnected graphs. The authors offer two methods of generalization of vanilla GCNN for disconnected graphs.

The main idea of the first method is constructing additional vector representations of molecules $\mathbf{h}_{m_i}^{(l)}$ and reaction $\mathbf{h}_r^{(l)}$. Vector representations of atoms in molecule is connected with corresponding molecule representation and the molecule representation connected with the reaction representation (see Fig. 3a). Constructing a new node in a molecular graph is not a novel approach. A similar mechanism was successfully applied to different tasks in computational drug development [46]. New updating rules are displayed in equations,

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

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

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

Where new bonds have a different type on different levels of vector representation.

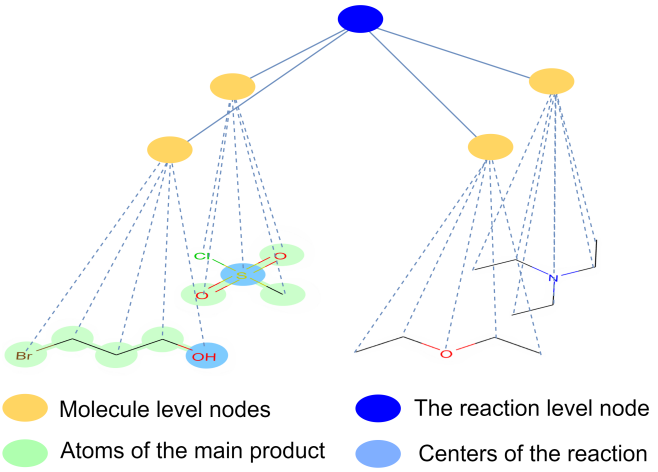
The second proposed approach uses attention-mechanism to aggregate information across nodes in a disconnected graph of source molecules. Attention-mechanism was originally proposed for the improvement of sequence-to-sequence models for the machine translation problem [47]. After that, the method was successfully applied to a variety of problems and was integrated into different neural architectures [48–50]. The output of the method depends on all inputs with trainable coefficients.

The authors offer using encoder of Transformer [51] (see Fig. 3b) for aggregation information across nodes. The core feature of the model is a multi-head attention. In particular, one-head attention is a self-attention. The mapping is mathematically formulated in equation,

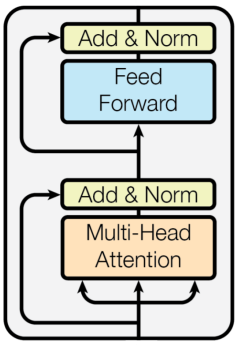
$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{K}\mathbf{Q}^\top}{\sqrt{d_{\text{model}}}}\right)\mathbf{V}.$$

Where $\mathbf{Q},\mathbf{K},\mathbf{V}$ are queries, keys, values; d_{model} is a dimension of key. A generalization of the formula for several heads is represented in equations,

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



(a) Extended molecular graph with introduced reaction and molecular level pseudo-nodes. The structure is suitable for applying GCNNs for disconnected graphs.



(b) Architecture of the Transformer Encoder layer.

Рис. 3: Two techniques that generalize GCNN on disconnected graphs. The figure 3a demonstrates the extended molecular graph, which unites source molecules. The figure 3b illustrates the encoder block of the Transformer architecture. Outputs of GCNN is processed with the Transformer to exchange information between graph components.

The model consists of several technical features which make the training process easier and more efficient: residual connections, norm layers, feed-forward layers [52, 53].

Construction of output and Loss function. The final vector representations are passed to a fully-connected neural network to get the probability of the node's class,

$$\mathbf{h}_i^{(l+1)} = \text{ReLU}\left(\text{linear}(\mathbf{h}_i^{(l)})\right),$$
$$\hat{\mathbb{P}}(y_i = 1) = \text{sigmoid}\left(\text{linear}(\mathbf{h}_i^{(n)})\right).$$

Where the final non-linearity is sigmoid function.
Value of loss function for a reaction is an average cross-entropy,

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{\mathbb{P}}(y_i = 1) + (1 - y_i)\log(1 - \hat{\mathbb{P}}(y_i = 1))\right).$$

General loss is an average of losses on each reaction.

Multi-task learning. Experiments demonstrate that learning multiple related tasks from data at the same time increases model performance comped with learning these tasks independently [54]. The authors focus on two related node classification problems in a disconnected graph of source molecules in a chemical reaction. Therefore, the multi-task learning technique was adapted to the classification problems. The proposed architecture consists of two fully-connected heads that construct output from vector representation of nodes for either classification problem. The final value of loss function is a sum of losses on both heads.

Experiment

The model was carefully selected with a step-by-step procedure(see Tab. 2). The procedure is an evolution of the model that display the importance of each proposed feature.

Modification\Model	BASE	EG	T	EGT	EGTB	EGTBF	MT_EGTBF
Extended molecular graph	-	+	-	+	+	+	+
Self-Attention	-	-	+	+	+	+	+
Types of bonds	-	-	-	-	+	+	+
Features of nodes from RDKit	-	-	-	-	-	+	+
Multi-task learning	-	-	-	-	-	-	+

Таблица 2: The set of models which demonstrates importance of each proposed modifications.

The first experiment shows that the architecture generalizes previously developed graph convolutional neural networks to disconnected graphs. The next experiments explain the importance of knowledge about source molecular graphs to achieve a high score. The last experiment illustrates that multi-task learning is a useful technique in solving problems. It reduces computational costs and increases quality when a model learn correlated tasks at the same time. Two atom classification problems are solved on USPTO_STEREO dataset (see Tab. 1) with one major constrain, the number of atoms in source molecules is less than 50.

The simplest model (**BASE**) consists of RGCNN and FCN parts and takes a disconnected graph of source molecules. Only types of nodes assume to be known. The model is unnatural for the atom classification problem in a disconnected graph. It does not use any problem-specific information and deny passing data between components in the disconnected graph. A comparison with the model displays that all the proposed modifications increase the model quality. The model demonstrates weak results (see Tab. 3) because of the final class of atom in a molecule depends only on atoms in the molecule. However, intermolecular interaction is the primary mechanism of chemical reactions. Using the extended molecular graph (**EG**) as an input of RGCNN prevents the model from the weakness. A significant increase of the results proves that additional representation of molecules and whole reaction simulates intermolecular interaction in the chemical reaction and exchanges information across source molecular graphs. Another proposed generalization of RGCNN is using Encoder of Transformer (**T**) architecture after convolution layers. The modification demonstrates better results than using an extended molecular graph. A combination of both changes (**EGT**) increases quality compared with each one.

Another quality of an excellent method for node classification in molecular graphs is using additional knowledge about atoms and bonds. The proposed method is suitable to add the features of edges and nodes naturally. Embeddings of nodes contain information about a variety of atom properties. Moreover, the relational structure of graph convolutional layers simulates different types of chemical bonds.

The next model (**EGTB**) works with different types of chemical bonds: single, double, triple, aromatic. The information results in a quality increase. The development corresponds to prior knowledge that type of chemical bonds impact the mechanism of

the reaction. The most valuable influence on the final result gives usage of different calculated properties of the atoms while initializing the embeddings of nodes in the extended molecular graph. The properties are degree, explicit valence, hybridization, implicit valence, aromaticity, implicitness, number of explicit hydrogenous, number of implicit hydrogenous, is a ring, number of radical electrons, formal charge. Adding the properties to the model (**EGTBF**) significantly improves model quality. The main product mapping quality rises to 61% full-match accuracy; centers of the reaction is detected with 60% full-match accuracy. The experiment displays that the model has an interpretable architecture that can take different properties of atoms and chemical bonds to show better performance on the dataset.

The considered classification problems are correlated. Prediction only atoms of the main product determine part of the centers of the reaction. Learning different correlated problems from data is a popular technique that increases model quality in a variety of problems. The authors apply the approach to the considered problems (**MT_EGTBF**). The modification slightly increases model quality in both cases. Moreover, the model is computationally efficient because it has shared RGCNN and Transformer parts and task-specific FCNNs.

Results

The authors construct a set of experiments that demonstrate the power of MolsNet. The final model achieves 61% (see Tab. 3) full-match accuracy on detection of the main product and 60% on detection centers in the reaction. Moreover, we demonstrate that all proposed modifications are significant and result in an increase of the final quality.

	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

Таблица 3: Results of the experiments. FM is an average full-match accuracy. F_1 is an average F_1-measure between ground-truth classes and predicted classes of atoms in a reaction.

All experiments were designed with PyTorch [55] and DGL [56] frameworks and run on Nvidia 1080Ti. Five epochs of learning the best architecture take approximately 4 hours.

Model analysis

As previously mentioned, there is a limitation of the number of source atoms in conducted experiments. The restriction reduces the used dataset by 25% and makes the model faster and smaller, which is proper for training multiple models. Learning the best model for higher limitation shows that model quality decrease slightly. The limit of 150 atoms reduces the dataset by less than 2%. Quality of the main product detection lowers by 2%, reaction centers by 3%. The model is robust for the length of the molecules.

On the next step, the authors investigate the dependency of the model quality on the length of source molecules and the number of centers in a reaction (see Fig. 5). Overall, the quality of the model does not depend on number of atoms in source molecules. The fact demonstrates that MolsNet has advantage above sequence-to-sequence approaches which quality often has a drop for long inputs. However, the quality dramatically decrease with increasing the number of centres. Multiple centers in a reaction mean that chemical compounds changes in the reaction significantly, and the process is complex(see Fig. 4). The analysis demonstrates that the model is more flexible for inputs of different lengths than sequence-to-sequence models.

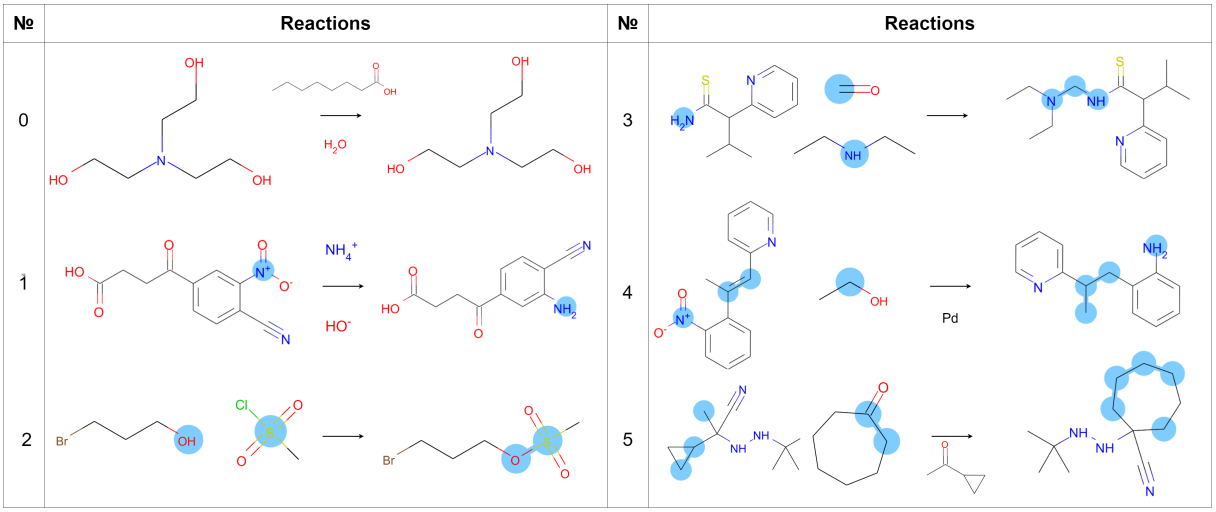


Рис. 4: Examples of reactions with different number of centres.

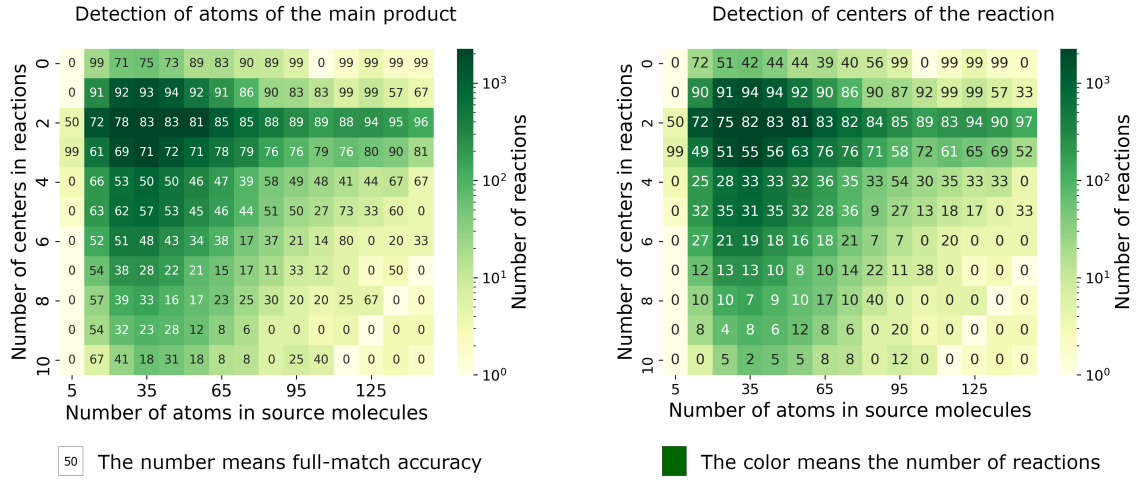


Рис. 5: The analysis of the dependency of the MT_EGTBF model quality on the number of atoms in source molecules and the number of centers. The color in the heatmaps illustrates the distribution of reactions in the test part of the USPTO_STEREO dataset. Annotated values are the percent of the right predictions in terms of full-match accuracy. The left figure demonstrates the quality of the main product mapping, the right figure displays the quality of detection of the centers

Interpretability of the model

The rest of the experiments are devoted to analyzing the proposed method. Firstly, the authors investigate vector representations of reactions. The best model (**MT_EGTBF**) demonstrates that pseudo-nodes in the extended graph of source molecules learn chemical information about the whole reaction. Similar representation correspond chemical reactions which have similar mechanism (see Fig. 6). The authors take the USPTO_50k dataset [35, 57], which contains 50 thousands of reactions of ten different classes to investigate the vector representations. The five largest and smallest classes are separately studied because the reactions in the dataset are very unbalanced. TSNE [58] maps (see. Fig. 7) show that space of reaction's representation contains information about class of reaction. The space is not perfect because the properties of the reaction representation space are learned unsupervised. However, the result demonstrates that the model has the potential to create high-quality descriptors of molecules, reactions, sets of molecules.

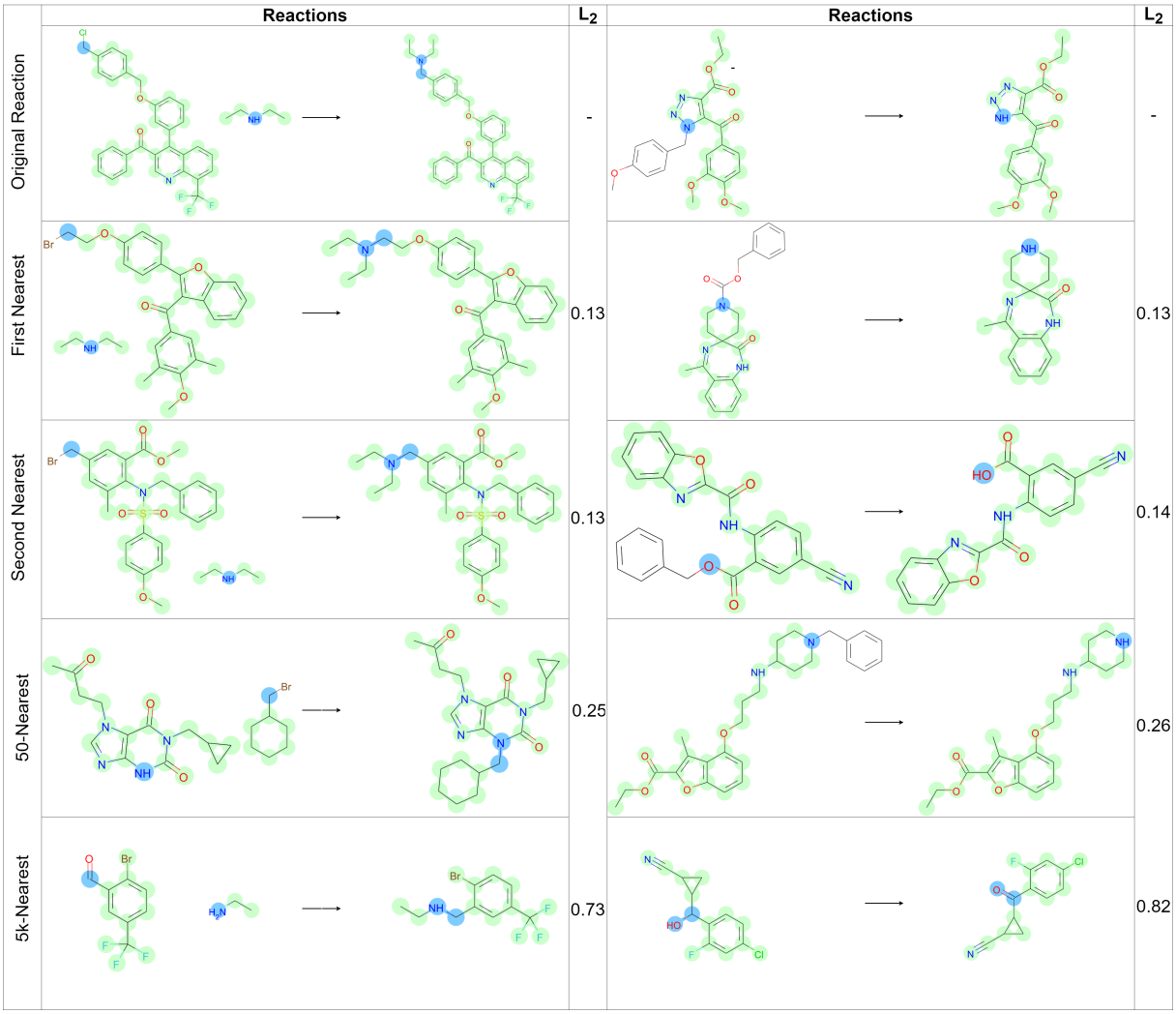
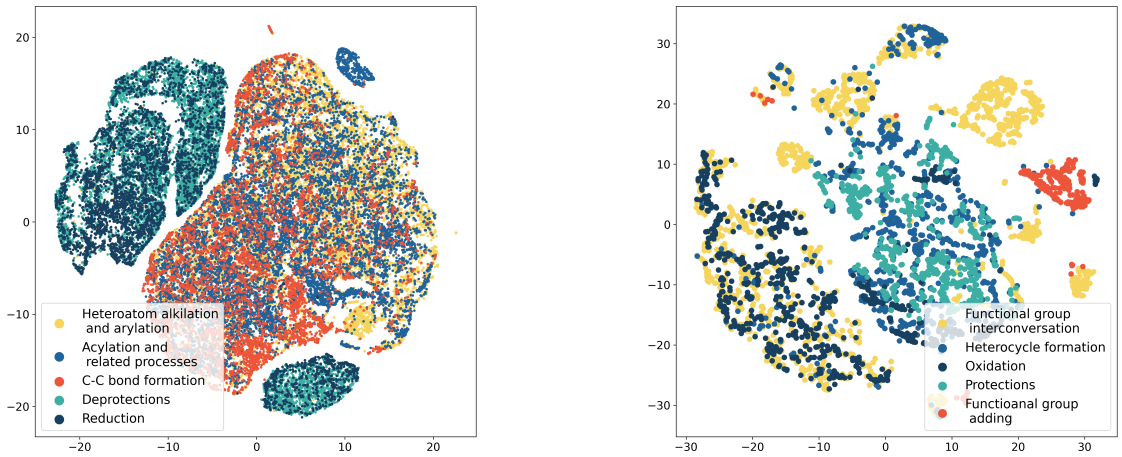


Рис. 6: Examples of nearest reactions. The figure shows that a similar vector representation of chemical reactions corresponds to reactions with the same mechanism.

Discussion

To sum up, the authors propose a novel architecture of neural network MolsNet which generalizes graph convolution neural network for a disconnected graph. Experiments demonstrate that the method demonstrates excellent results on the atom mapping problem. Moreover, the proposed model effectively uses knowledge about the structure of given molecular graphs. Additional information about chemical bond types between atoms and properties of these atoms significantly improves the model performance. Solving both node classification problems at the same time results in a slight increase in the quantity of the solution. The technique makes the model computationally efficient.

The authors developed an accurate and efficient method for the atom mapping and the main product prediction in the chemical reaction. The problem is solved with the MolsNet neural network. The model was analyzed on the large-scale USPTO_STEREO dataset. The set of experiments demonstrates that the model is capable of process chemical information about molecular graph structure included numerical characteristics of atoms and bonds in source molecules. Analysis of the final model display flexibility of the model



(a) (b)

Рис. 7: The TSNE maps of vector representations of reactions are here. Colors correspond to classes of chemical reactions in the USPTO_50k dataset. The figure 7a displays reactions from five most frequent classes. The reactions make up 90% of USPTO_50k dataset. The figure 7b represents five less frequent classes.

to molecule's size, learning chemical insights from given reactions. However, the model has several limitations. The proposed architecture is not suitable for multiple mappings detection. The quality of the solution drops dramatically with increasing reaction number of centers in the reaction. A large number of centers means the complexity of the reaction. This paper considers only application on MolsNet to molecular graphs in chemical reactions, although the approach presented can be applied to disconnected graphs in general. It expands the GCNNs for various problems in computational chemistry such as atom classification in molecular graphs, classification of molecular graphs, different prediction of atom's properties in reactions and solutions. Using local features of source molecular graphs increases the value of the method for usage in complex systems. In future work, the authors plan to create a software pipeline that makes the proposed methodology applicable in industry.

References

[1] Machine learning for molecular and materials science / Keith T Butler, Daniel W Davies, Hugh Cartwright et al. // Nature. — 2018. — Vol. 559, no. 7715. — P. 547.

[2] Survey of machine learning techniques in drug discovery / Natalie Stephenson, Emily Shane, Jessica Chase et al. // Current drug metabolism. — 2019. — Vol. 20, no. 3. — P. 185–193.

[3] A survey on Big Data and Machine Learning for Chemistry / Jose F Rodrigues Jr, Larisa Florea, Maria CF de Oliveira et al. // arXiv preprint arXiv:1904.10370. — 2019.

[4] Advances and Perspectives in Applying Deep Learning for Drug Design and Discovery / Celio Lipinski, Vinicius Maltarollo, Patricia Oliveira et al. // Frontiers in Robotics and AI. — 2019. — Vol. 6. — P. 108.

[5] Segler Marwin HS, Preuss Mike, Waller Mark P. Planning chemical syntheses with deep neural networks and symbolic AI // Nature. — 2018. — Vol. 555, no. 7698. — P. 604.

[6] "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models / Philippe Schwaller, Theophile Gaudin, David Lanyi et al. // Chemical science. — 2018. — Vol. 9, no. 28. — P. 6091–6098.

[7] Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation / Philippe Schwaller, Teodoro Laino, Théophile Gaudin et al. // arXiv preprint arXiv:1811.02633. — 2018.

[8] A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity / Connor W Coley, Wengong Jin, Luke Rogers et al. — 2018.

[9] Corey EJ, Wipke W Todd. Computer-assisted design of complex organic syntheses // Science. — 1969. — Vol. 166, no. 3902. — P. 178–192.

[10] Cerebral Aneurysm Multicenter European Onyx (CAMEO) trial: results of a prospective observational study in 20 European centers / Andrew J Molyneux, Saruhan Cekirge, Isil Saatci, Gyula Gál // American Journal of Neuroradiology. — 2004. — Vol. 25, no. 1. — P. 39–51.

[11] A new treatment of chemical reactivity: Development of EROS, an expert system for reaction prediction and synthesis design / Johann Gasteiger, Michael G Hutchings, Bernd Christoph et al. // Organic Synthesis, Reactions and Mechanisms. — Springer, 1987. — P. 19–73.

[12] Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry / Ivar Ugi, Johannes Bauer, Klemens Bley et al. // Angewandte Chemie International Edition in English. — 1993. — Vol. 32, no. 2. — P. 201–227.

[13] Satoh Hiroko, Funatsu Kimito. SOPHIA, a knowledge base-guided reaction prediction system-utilization of a knowledge base derived from a reaction database // Journal of chemical information and computer sciences. — 1995. — Vol. 35, no. 1. — P. 34–44.

[14] Nikitin Filipp, Strijov Vadim. Graph neural network learning for chemical compounds synthesis // Mathematical Methods of Pattern Recognition conference (MMPR-19) / Russian Academy of Sciences. — 2019. — P. 311–312.

[15] Sullivan Thomas. A Tough Road: Cost to develop one new drug Is $2.6 billion; approval rate for drugs entering clinical development is less than 12% // Policy & Medicine. — 2019.

[16] Principles of early drug discovery / James P Hughes, Stephen Rees, S Barrett Kalindjian, Karen L Philpott // British journal of pharmacology. — 2011. — Vol. 162, no. 6. — P. 1239–1249.

[17] Modeling relational data with graph convolutional networks / Michael Schlichtkrull, Thomas N Kipf, Peter Bloem et al. // European Semantic Web Conference / Springer. — 2018. — P. 593–607.

[18] Schreiber Stuart L. Target-oriented and diversity-oriented organic synthesis in drug discovery // Science. — 2000. — Vol. 287, no. 5460. — P. 1964–1969.

[19] Schneider Gisbert. Automating drug discovery // Nature Reviews Drug Discovery. — 2018. — Vol. 17, no. 2. — P. 97.

[20] Schneider Gisbert, Fechner Uli. Computer-based de novo design of drug-like molecules // Nature Reviews Drug Discovery. — 2005. — Vol. 4, no. 8. — P. 649.

[21] Computational methods in drug discovery / Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, Edward W Lowe // Pharmacological reviews. — 2014. — Vol. 66, no. 1. — P. 334–395.

[22] Recent trends in deep learning based natural language processing / Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria // ieee Computational intelligenCe magazine. — 2018. — Vol. 13, no. 3. — P. 55–75.

[23] Yu Dong, Deng Li. Deep learning and its applications to signal and information processing [exploratory dsp] // IEEE Signal Processing Magazine. — 2010. — Vol. 28, no. 1. — P. 145–154.

[24] Deep learning for computer vision: A brief review / Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis // Computational intelligence and neuroscience. — 2018. — Vol. 2018.

[25] Weininger David. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules // Journal of chemical information and computer sciences. — 1988. — Vol. 28, no. 1. — P. 31–36.

[26] Molecular de-novo design through deep reinforcement learning / Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, Hongming Chen // Journal of cheminformatics. — 2017. — Vol. 9, no. 1. — P. 48.

[27] Automatic chemical design using a data-driven continuous representation of molecules / Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud et al. // ACS central science. — 2018. — Vol. 4, no. 2. — P. 268–276.

[28] Syntax-directed variational autoencoder for structured data / Hanjun Dai, Yingtao Tian, Bo Dai et al. // arXiv preprint arXiv:1802.08786. — 2018.

[29] De Cao Nicola, Kipf Thomas. MolGAN: An implicit generative model for small molecular graphs // arXiv preprint arXiv:1805.11973. — 2018.

[30] Convolutional networks on graphs for learning molecular fingerprints / David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre et al. // Advances in neural information processing systems. — 2015. — P. 2224–2232.

[31] Organic synthesis provides opportunities to transform drug discovery / David C Blakemore, Luis Castro, Ian Churcher et al. // Nature chemistry. — 2018. — Vol. 10, no. 4. — P. 383.

[32] Ojima Iwao. Catalytic asymmetric synthesis. — John Wiley & Sons, 2004.

[33] De Meijere Armin, Bräse Stefan, Oestreich Martin. Metal catalyzed cross-coupling reactions and more, 3 Volume Set. — John Wiley & Sons, 2013.

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

[35] Retrosynthetic reaction prediction using neural sequence-to-sequence models / Bowen Liu, Bharath Ramsundar, Prasad Kawthekar et al. // ACS central science. — 2017. — Vol. 3, no. 10. — P. 1103–1113.

[36] Huczko Appl. Template-based synthesis of nanomaterials // Applied Physics A. — 2000. — Vol. 70, no. 4. — P. 365–376.

[37] Atom mapping with constraint programming / Martin Mann, Feras Nahar, Norah Schnorr et al. // Algorithms for Molecular Biology. — 2014. — Vol. 9, no. 1. — P. 23.

[38] Fooshee David, Andronico Alessio, Baldi Pierre. ReactionMap: An efficient atom-mapping algorithm for chemical reactions // Journal of chemical information and modeling. — 2013. — Vol. 53, no. 11. — P. 2812–2819.

[39] Accurate atom-mapping computation for biochemical reactions / Mario Latendresse, Jeremiah P Malerich, Mike Travers, Peter D Karp // Journal of chemical information and modeling. — 2012. — Vol. 52, no. 11. — P. 2970–2982.

[40] Automatic mapping of atoms across both simple and complex chemical reactions / Wojciech Jaworski, Sara Szymkuć, Barbara Mikulak-Klucznik et al. // Nature communications. — 2019. — Vol. 10, no. 1. — P. 1–11.

[41] Kipf Thomas N, Welling Max. Semi-supervised classification with graph convolutional networks // arXiv preprint arXiv:1609.02907. — 2016.

[42] Lowe Daniel Mark. Extraction of chemical structures and reactions from the literature : Ph. D. thesis / Daniel Mark Lowe ; University of Cambridge. — 2012.

[43] Edwards Ward, Barron F Hutton. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement // Organizational behavior and human decision processes. — 1994. — Vol. 60, no. 3. — P. 306–325.

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

[45] A comprehensive survey on graph neural networks / Zonghan Wu, Shirui Pan, Fengwen Chen et al. // arXiv preprint arXiv:1901.00596. — 2019.

[46] Li Junying, Cai Deng, He Xiaofei. Learning graph-level representation for drug discovery // arXiv preprint arXiv:1709.03741. — 2017.

[47] Luong Minh-Thang, Pham Hieu, Manning Christopher D. Effective approaches to attention-based neural machine translation // arXiv preprint arXiv:1508.04025. — 2015.

[48] Hu Dichao. An introductory survey on attention mechanisms in NLP problems // Proceedings of SAI Intelligent Systems Conference / Springer. — 2019. — P. 432–448.

[49] Graph attention networks / Petar Veličković, Guillem Cucurull, Arantxa Casanova et al. // arXiv preprint arXiv:1710.10903. — 2017.

[50] Bottom-up and top-down attention for image captioning and visual question answering / Peter Anderson, Xiaodong He, Chris Buehler et al. // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 6077–6086.

[51] Attention is all you need / Ashish Vaswani, Noam Shazeer, Niki Parmar et al. // Advances in neural information processing systems. — 2017. — P. 5998–6008.

[52] Inception-v4, inception-resnet and the impact of residual connections on learning / Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alexander A Alemi // Thirty-First AAAI Conference on Artificial Intelligence. — 2017.

[53] Ba Jimmy Lei, Kiros Jamie Ryan, Hinton Geoffrey E. Layer normalization // arXiv preprint arXiv:1607.06450. — 2016.

[54] Evgeniou Theodoros, Pontil Massimiliano. Regularized multi–task learning // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — 2004. — P. 109–117.

[55] PyTorch: An imperative style, high-performance deep learning library / Adam Paszke, Sam Gross, Francisco Massa et al. // Advances in Neural Information Processing Systems. — 2019. — P. 8024–8035.

[56] Deep graph library: Towards efficient and scalable deep learning on graphs / Minjie Wang, Lingfan Yu, Da Zheng et al. // arXiv preprint arXiv:1909.01315. — 2019.

[57] Schneider Nadine, Stiefl Nikolaus, Landrum Gregory A. What's what: The (nearly) definitive guide to reaction role assignment // Journal of chemical information and modeling. — 2016. — Vol. 56, no. 12. — P. 2336–2346.

[58] Maaten Laurens van der, Hinton Geoffrey. Visualizing data using t-SNE // Journal of machine learning research. — 2008. — Vol. 9, no. Nov. — P. 2579–2605.