DRACON: Disconnected Graph Neural Network for Atom Mapping in Chemical Reactions

Filipp Nikitin,[*,†] Olexandr Isayev,[*,‡] and Vadim Strijov[*,†]

†*Moscow Institute of Physics and Technology*

‡*Department of Chemistry, Carnegie Mellon University*

E-mail: filipp.nikitin@phystech.edu; olexandr@olexandrisayev.com; strijov@phystech.edu

Abstract

Machine learning solved many challenging problems in computer-assisted synthesis prediction (CASP). We formulate a reaction prediction problem in terms of node-classification in a disconnected graph of source molecules and generalize a graph convolution neural network for disconnected graphs. Here we demonstrate that our approach can successfully predict reaction outcome and atom-mapping during a chemical transformation. A set of experiments using the USPTO dataset demonstrates excellent performance and interpretability of the proposed model. Implicitly learned latent vector representation of chemical reactions strongly correlates with the class of the chemical reaction. Reactions with similar templates group together in the latent vector space.

Introduction

Drug discovery is 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. Advances in computer science and machine learning have changed how the drug discovery process is

performed.[1–4] Recent works aimed at the prediction of new chemical compounds with the desired profile of properties, including efficacy, pharmacokinetics, and safety.[1,5,6] This is a computationally hard problem because the space of available molecules is huge.[7] Machine learning methods make this problem tractable.[8–11] Generative models for molecules were built with a recurrent neural network that generates SMILES[12] representation of target compound character by character.[6,13,14] Variational autoencoders and graph neural networks were successfully applied to the problem.[9,10,15] The generative algorithms explore chemical space beyond the currently enumerated libraries. Therefore, we rarely know *a priori* how these molecules could be synthesized.

Today, CASP[16] is a particularly active field of chemical research. Organic chemists recognized the potential of computational methods in practice and developed the first rule-based method (OCSS) 50 years ago.[17] Similar works are CAMEO,[18] EROS,[19] IGOR,[20] SOPHIA.[21] Medical chemists use a huge set of unstructured rules to predict products in the reaction. Computer-aided retrosynthesis would be a valuable tool, but at present, it is slow and provides results of unsatisfactory quality.[22]

Modern approaches to the problem rely on deep learning methods. Neural network architectures for Neural Machine Translation[23] were adapted to the forward synthesis problem.[22,24,25] These methods use SMILES representation of the reagents, reactants, and products. It translates the source string to the product string character by character. RNNs and Transformer[26] architectures for NMT demonstrate the excellent performance of the outcome prediction problem. The Transformer architecture is now the state of the art solution for many tasks. These sequence-to-sequence methods have several disadvantages. They do not take advantage of the graph structure of molecules, SMILES language construction and depth of chemical knowledge. Therefore, graph convolution neural networks (GCNNs) were proposed to evaluate the probability of a bond between two nodes.[8]

The sequence-to-sequence models use SMILES strings of reagents and products.[27] More sophisticated methods require the atom mapping.[28] The atom mapping of a chemical reac-

tion 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, facilitating substructure searches, identifying metabolic pathways.[29–31] Labeling atom mapping requires tedious manual annotation by human domain experts. Most of the reactions in databases are not mass-balanced and not atom-mapped. It creates problems for automated machine understanding of chemical reactions.[32]

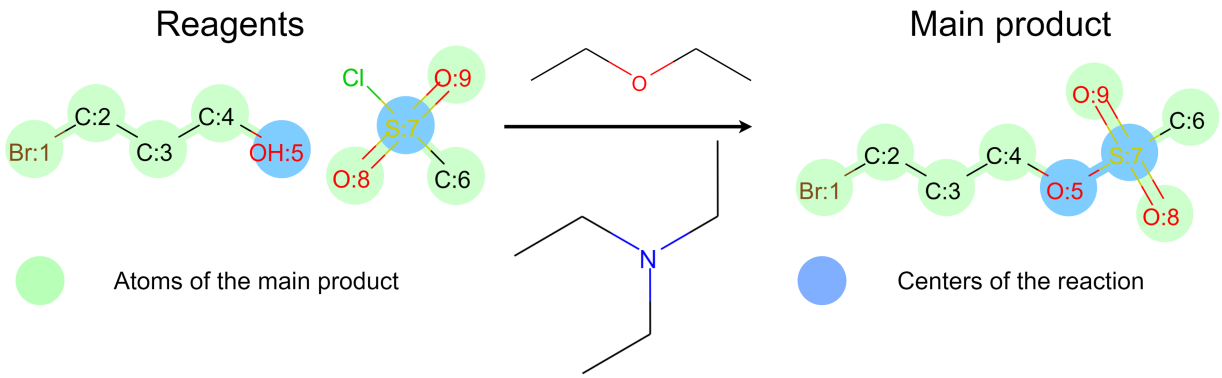


Figure 1: A chemical reaction maps reagents into products. On the molecules of reagents, two types of atoms are labeled: atoms of the main product and centers of the chemical reaction. Centers are the atoms that change their characteristics.

We propose a model that affords both predicting reaction 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 reaction centers are found. Centers of the reaction are atoms of the main product. The atoms change their configuration during the reaction. The configuration of an atom is an aggregate of characteristics of the atom and adjacent bonds. In terms of graph theory, both tasks are node-classification in a disconnected graph of reactant and product molecules. The novel **D**isconnected G**r**aph **A**ttention **Con**volution neural network (DRACON) solves the node-classification tasks. Atoms of the main product and centers of the reaction determine the outcome in the majority of reactions.

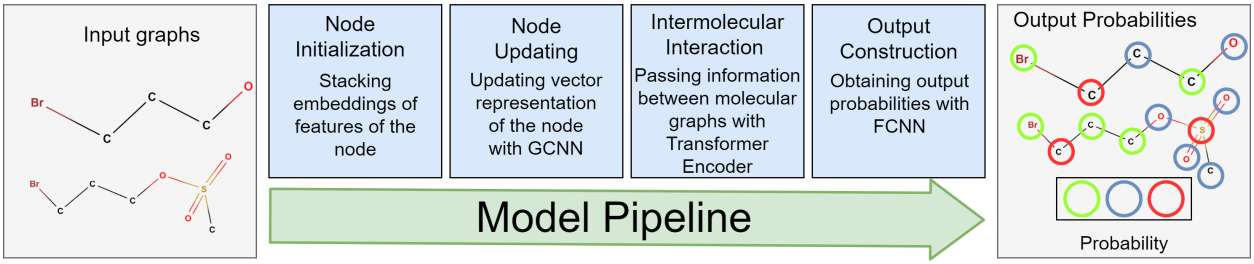


Figure 2: The DRACON architecture. Each step in this pipeline naturally corresponds to the structure of given molecular graphs. It uses different local features of atoms in molecules to construct an initial representation of nodes. The final atom representation is given according to the adjacent node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

Methods

The overall model pipeline consists of four blocks (see Fig. 2). Firstly, each atom is mapped to a real vector according to its characteristics in the molecule. The model uses numerical characteristics of atoms. Secondly, the vectors are updated with Relational Graph Convolution Neural Network (RGCNN).[33] The RGCNN generalizes Graph Convolution Neural Network[34] for graphs with different edge types that correspond to chemical bonds. In this work, we use extended molecular graphs with molecules' and reaction's level nodes to pass information across different molecules.[11] 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 works, DRACON has several novel aspects in terms of neural network architecture. DRACON generalizes the graph convolution neural network for the disconnected molecular graphs. The natural structure of the DRACON is suitable to add features of molecules, atoms like valencies, hybridization, types of chemical bonds, etc.

Experiments conducted on the dataset of reactions, which were extracted from the US patents (USPTO). The model demonstrates excellent performance in both node-classification tasks. It is a generalization of RGCNN architecture for disconnected graphs. DRACON uses an unsupervised approach to atom-mapping and bridges the gap between data-driven approaches and traditional rule-based systems. Finally, the model analysis illustrates that it gains substantial chemical insight, and one could differentiate and group chemical reactions by their types in a fully unsupervised fashion.

Problem statement. Molecules in a chemical reaction can be considered as a disconnected graph $\mathbf{G}$ with features of nodes and edges. Our model $\mathcal{F}$ maps this graph $\mathbf{G}$ to the labels $\hat{\mathbf{y}}$ of its nodes, $\mathbf{y}$ are ground-truth labels. Let $\mathcal{L}$ is an internal criterion of the model quality. Then the problem of model selection for node classification in a disconnected graph is formulated,

$$\mathcal{F} : \mathbf{G} \to \hat{\mathbf{y}}$$

$$\min_{\mathcal{F}} \mathcal{L}(\mathbf{y}, \hat{\mathbf{y}}).$$

Initialization of vector representation for nodes. The description of the atom consists of K categorical features. Each feature of the atom is represented as boolean vector $\mathbf{b}_{ik}$ with one-hot encoding.[35] The vector embeds into real value space with multiplying by the real value weight matrix $\mathbf{W}_k$.

$$\mathbf{h}_{ik}^{(0)} = \mathbf{W}_k \mathbf{b}_{ik}.$$

Where i is an atom, k is an index of the feature, and a higher index shows an index of the layer in the entire neural network.

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

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

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,[34] which is often the number of input edges.

One disadvantage of the model (1) is the assumption that edges in the graph are the same. The type of chemical bond is an important feature of a molecular graph. In Relational Graph Convolution Neural network (RGCNN[33]), 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 r, there is a unique weight matrix $\mathbf{W}_r^{(l)}$.

Passing information between graph components. The core of the proposed model is Graph Neural Network.[36] The output of the model must depend on all atom's representation in source molecules. Therefore, the updating mechanism of RGCNN must be generalized to work with disconnected graphs. The authors offer two methods of generalization of original 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 $\mathbf{h}_i^{(l)}$ is connected

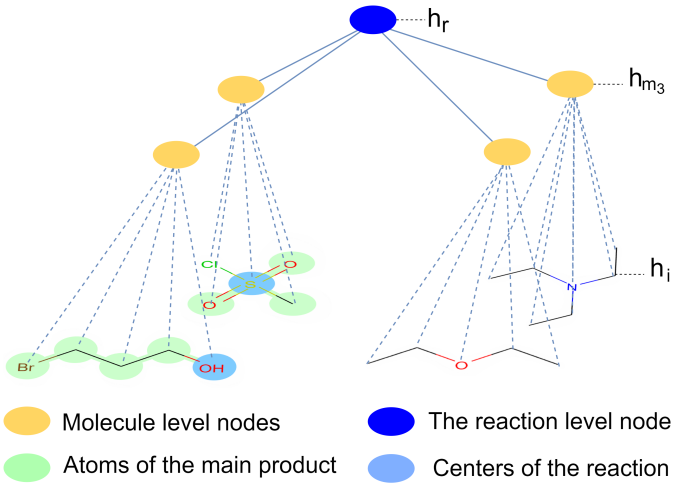
with corresponding molecule representation $\mathbf{h}_{m_i}^{(l)}$ and the molecule representation connected with the reaction representation $\mathbf{h}_r^{(l)}$ (see Fig. 3a). Modified updating rules are displayed in equations,

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

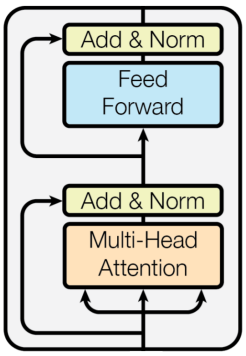
$$\mathbf{h}_{m_k}^{(l+1)} = \text{ReLU}\left(\mathbf{W}^{(l)}\mathbf{h}_{m_k}^{(l)} + \mathbf{W}_{rl}^{(l)}\mathbf{h}_r^{(l)} + \sum_{j\in m_k}\frac{1}{|m_k|}\mathbf{W}_{ml}^{(l)}\mathbf{h}_j^{(l)}\right),$$

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

Where $|m_k|$ is a number of atoms in a molecule, $|M|$ is a number of molecules in a reaction.



(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 (The image from original paper[26]).

Figure 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 second proposed method uses an attention mechanism to aggregate information across nodes in a disconnected graph of source molecules.[26,37] The output of the method

depends on all inputs with trainable coefficients. In this work we propose using encoder of Transformer[26] (see Fig. 3b) for aggregation information across nodes. The core feature of the model is multi-head attention.

Multi-head attention transforms matrix of vector representation of nodes according to the equation,

$$\begin{aligned} \mathbf{H}^{(l+1)} &= \text{concat}[\text{head}_1, \text{head}_2, \ldots, \text{head}_h]\mathbf{W}^O, \\ \text{head}_i &= \text{Attention}\left(\mathbf{H}^{(l)}\mathbf{W}_i^Q, \mathbf{H}^{(l)}\mathbf{W}_i^K, \mathbf{H}^{(l)}\mathbf{W}_i^V\right). \end{aligned} \tag{2}$$

In the equation (2), Attention is a matrix function,

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

Where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are queries, keys, values; d_{model} is a dimension of key.

The model also takes advantage of several technical tricks that make the training process faster and more efficient: residual connections, normalization layers, feed-forward layers.[38,39]

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

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

Where the final non-linearity is a sigmoid function.

Loss function. Value of loss function for a reaction is an average cross-entropy,

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{\mathsf{P}}(y_i = 1) + (1 - y_i)\log(1 - \hat{\mathsf{P}}(y_i = 0))\right).$$

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.[40] We solve two correlated node classification problems. The multi-task learning technique was adapted to the problems with sharing parameters in the first three blocks of our model. The final value of loss function is a sum of losses for either classification problem.

Table 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[41]

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

Dataset of chemical reactions. 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.[41] The original patent information describes a complex chemical synthesis process consisting of multiple steps. The information summarised to a SMARTS[42] 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, [8,24] 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 300 thousands of reactions. The USPTO_50k contains 50 thousands of reactions, which has one of ten classes.

In the experiments, we use the USPTO_STEREO dataset with the original split into train, test, and validation parts. The USPTO_50k was used to analyze the model's insights.

The SMARTS representation of a reaction is converted to a molecular graph with open-source library RDKit.[43] The library is used to calculate atom features: degree, hybridization, aromaticity, implicitness, is a ring, number of radical electrons, formal charge.

Results and discussion

The model evaluation. The final model achieves 61% (see Table 3) full-match accuracy on detection of the main product and 60% on detection centers in the reaction. We can not compare the result with the previous works[8,25,32] because our research focuses on finding atoms of the main product and centers of reactions. We have not built a full solution for the entire atom mapping problem or outcome prediction problem. The model was carefully selected from the sequence of models of increasing expressivity (see Table 2). The selection illustrates the importance of each proposed modification.

Table 2: Consecutive hypothesis testing during model design.

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

The simplest model (**BASE**) consists of RGCNN and FCN blocks(see Fig. 2). It takes a disconnected graph of source molecules where only types of nodes assume to be known. The model does not use bond types and features of atoms and updates vector representation of atoms only inside a component of a disconnected graph. Therefore, the **BASE** model

Table 3: Results of the experiments. FM is an average full-match accuracy: the ratio of reaction in which all atoms classified correctly. F_1 is an average F_1-measure between ground-truth classes of atoms in the reaction and predicted classes.

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

demonstrates poor performance because the atom class's final class depends only on the local atomic environment in a single molecule. Most chemical reactions involve several reactants, catalysts, and solvents. Therefore intermolecular interactions are of primary importance for the mechanism. Using the extended molecular graph (**EG**) as an input of RGCNN expands the receptive field to the whole disconnected graph of all reactants. A significant increase in the model performance proves that introduced representation of molecules and whole reaction mimics this chemistry. Another proposed generalization of RGCNN is using Encoder of Transformer (**T**) architecture after convolution layers. The modification boosts the model quality compared with BASE, but EG outperforms the T model. Both modifications (**EGT**) makes the results better compared to EG and T individually.

The next model (**EGTB**) introduces different types of chemical bonds: single, double, triple, and aromatic. The information results in yet another performance increase. **EGTB** models encode prior knowledge that the type of chemical bonds impacts the mechanism of the reaction. Finally, using different RDKit computed properties of the atoms while initializing the embeddings of nodes in the extended molecular graph results in another performance boost. The properties are connection degree, hybridization, aromaticity, 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 59% full-match accuracy;

centers of the reaction is detected with 60% full-match accuracy. The experiment displays that the model has an intuitive chemically-interpretable architecture that can take advantage of different characteristics of atoms and chemical bonds.

The aforementioned models were trained separately for each task. However, prediction of atoms in the main product also determines part of the centers of the reaction. Joint learning of different correlated problems from data is a popular technique that increases model quality in a variety of problems.[40,44,45] We applied multi-task learning to our model (**MT_EGTBF**). The modification slightly increases model quality in both cases. Moreover, the model is computationally efficient because it shares RGCNN and Transformer parts and task-specific FCNNs. Described experiments were conducted on reactions with less than 50 atoms in source molecules to accelerate multiple experiments. Learning the best model for higher limitation shows that model quality decrease slightly.

Error analysis. We 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 the number of atoms in source molecules. However, the quality dramatically decreases with increasing the number of centers. Multiple centers in a reaction hint to a complex reaction mechanism, where chemical compounds undergo multiple transformation steps (see Fig. 4). Such reaction center analysis could be useful for detection of questionable reactions, like transformations with zero centers.

Chemical insights. Model interpretation is a significant component in any ML study.[46] In this section, we demonstrate how DRACON learns and memorizes useful information from disconnected reaction graphs. First, we investigate the learned latent vector representations of reactions. The best model (**MT_EGTBF**) demonstrates that pseudo-nodes in the extended graph (see Fig. 3a) of source molecules learn chemical information about the whole reaction. Looking for close neighbors in this space, we see that they share the same mechanism. In Fig. 6 we list two examples with corresponding L2 distances. As the distance

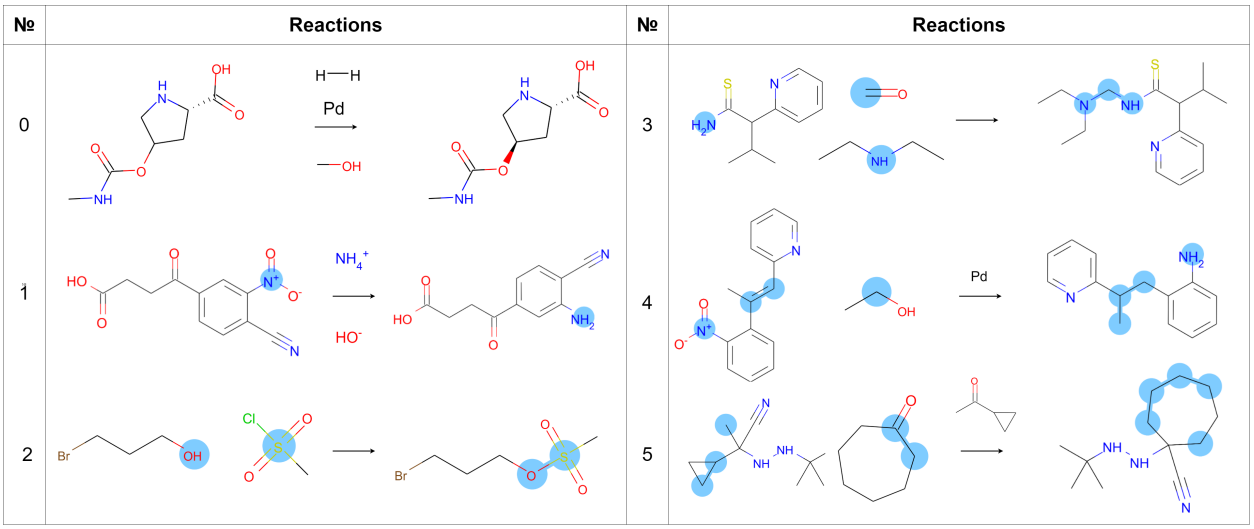


Figure 4: Examples of reactions with different number of centres.

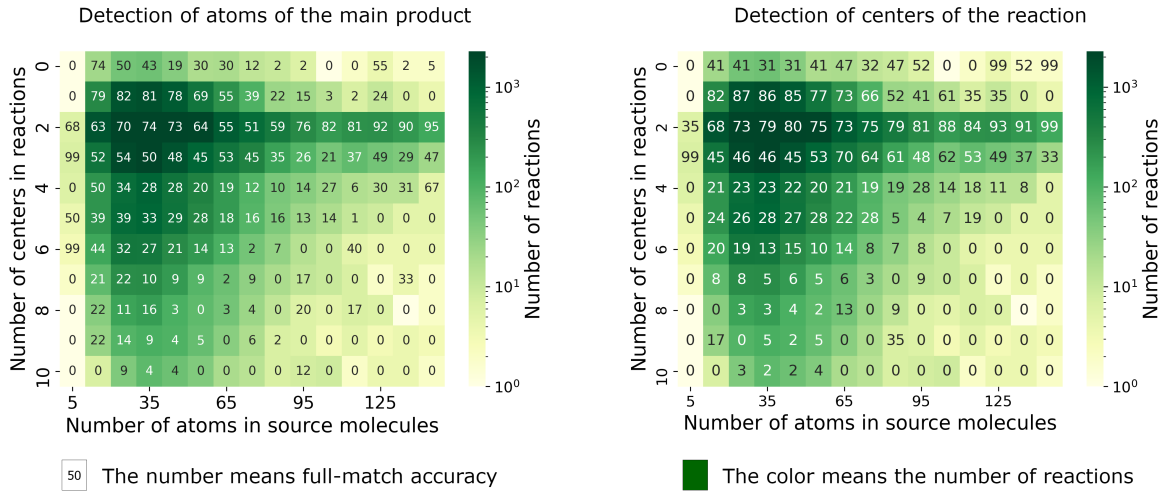


Figure 5: The analysis of the dependency of the MT_EGTBF model quality on the number of atoms in source molecules and the number of centers. The color in the heatmaps illustrates the distribution of reactions in the test part of the USPTO_STEREO dataset. Annotated values are the percent of the right 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

grows, the similarity drops, by the 50-th neighbor, you start to encounter very different reactions.

In other to visualize the chemical space of reactions, we selected the USPTO_50k dataset,[27,47] which contains 50 thousands of reactions that were annotated to ten different classes. Un-

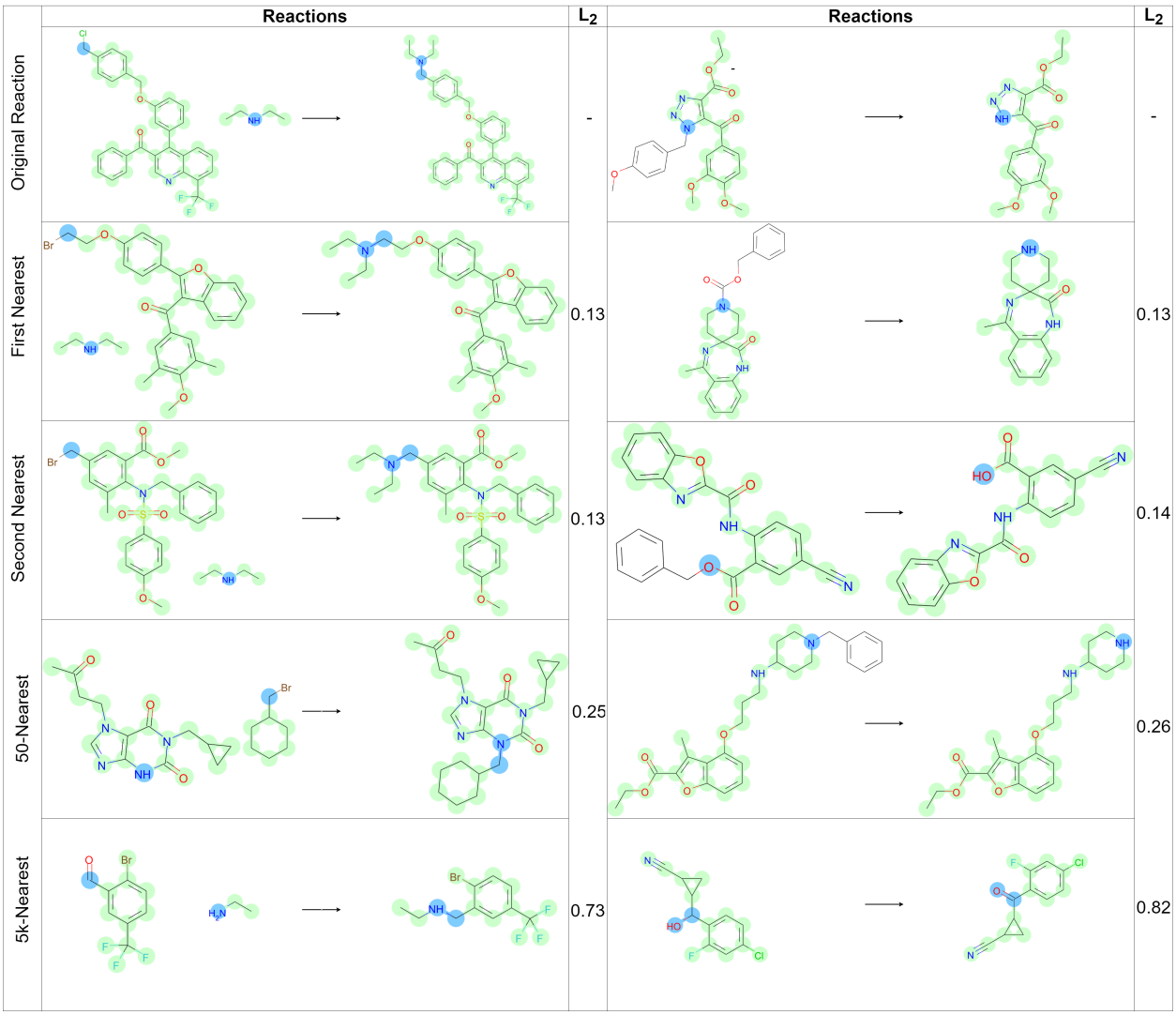
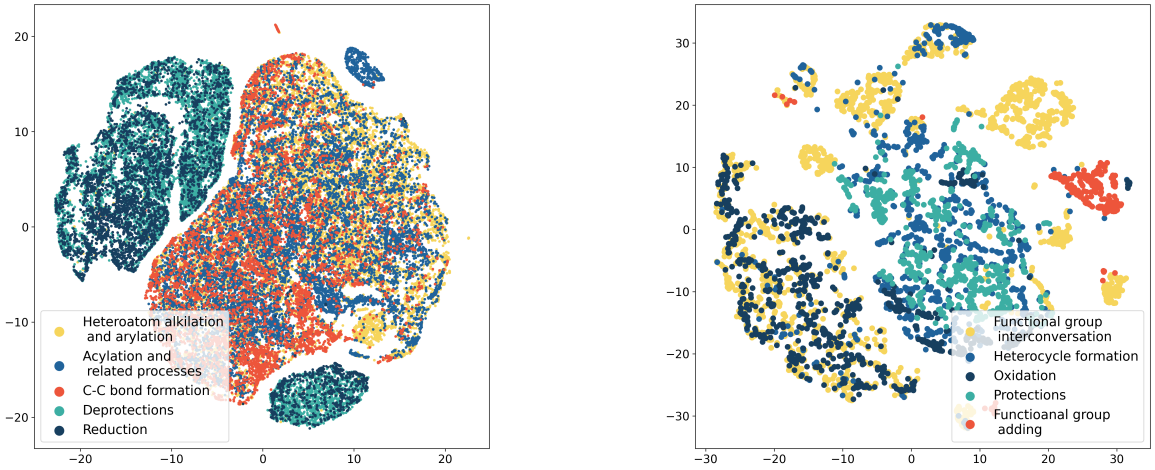


Figure 6: Examples of nearest reactions. The figure shows that a similar vector representation of chemical reactions corresponds to reactions with the same mechanism.

fortunately, this dataset is highly imbalanced. We separately visualized the five largest and smallest classes for clarity. TSNE[48] maps (see. Fig. 7) show significant correlation of reaction's representation with the class of reaction. The separation into classes is not perfect because the properties of the reaction representation space were learned fully unsupervised. However, the result demonstrates that the model has the potential to create high-quality descriptors for molecules, reactions, molecular sets.



(a) (b)

Figure 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.

Conclusions

We developed an interpretable and accurate model for outcome prediction and atom mapping in chemical reactions. A novel neural network architecture, DRACON were proposed for node classification problem in disconnected graphs. It generalizes graph convolution neural network for a disconnected graph with self-attention mechanism and learning hierarchical representations of source graphs. The model also generalizes the idea of graph representation learning with pseudo-nodes, which is state-of-the-art for a variety of problems in drug discovery.[11] The model was analyzed on the large-scale USPTO_STEREO dataset. The results demonstrate that DRACON predicts atoms of the main product and centers in a chemical reaction with high accuracy.

DRACON has an interpretable structure: it uses types of chemical bonds and characteristics of atoms in source molecules. The introduced pseudo-nodes of chemical reactions (see Fig. 3a) implicitly learn the similarity of chemical reactions.

This paper considers the application of DRACON to molecular graphs in chemical reactions. The approach presented can be applied to disconnected graphs in general. It expands the graph convolution neural networks 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 accuracy of the model for more complex tasks.

In this paper, we focus on finding atoms of the main product and centers of reaction. We plan to extend the methods for entire atom mapping and outcome prediction in chemical reactions in the future work. The model has several limitations. The proposed architecture is not suitable for multiple mappings detection like equivalent symmetric atoms in the molecule.

Conflicts of interest

There are no conflicts to declare.

Acknowledgement

O.I. acknowledges support from NSF CHE-1802789. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE)[49] award DMR110088, which is supported by NSF grant number ACI-1053575. We gratefully acknowledge the support from NVIDIA Corporation and express our special gratitude to Jonathan Lefman.

Supporting Information Available

Reference code implementation, trained models, and interactive visualization of reaction chemical space are freely available on GitHub: `https://github.com/isayevlab/DRACON`.

References

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547.

(2) Schneider, G. Automating drug discovery. *Nature Reviews Drug Discovery* **2018**, *17*, 97.

(3) Smith, J. S.; Roitberg, A. E.; Isayev, O. Transforming Computational Drug Discovery with Machine Learning and AI. *ACS Medicinal Chemistry Letters* **2018**, *9*, 1065–1069.

(4) Walters, W. P.; Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nature Biotechnology* **2020**, *38*, 143–145.

(5) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery* **2005**, *4*, 649.

(6) Popova, M.; Isayev, O.; Tropsha, A. Deep reinforcement learning for de novo drug design. *Science advances* **2018**, *4*, eaap7885.

(7) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational methods in drug discovery. *Pharmacological reviews* **2014**, *66*, 334–395.

(8) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Green, W.; Barzilay, R.; Jensen, K., et al. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. **2018**,

(9) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786* **2018**,

(10) De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973* **2018**,

(11) Li, J.; Cai, D.; He, X. Learning graph-level representation for drug discovery. *arXiv preprint arXiv:1709.03741* **2017**,

(12) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(13) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics* **2017**, *9*, 48.

(14) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **2018**, *4*, 268–276.

(15) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. Advances in neural information processing systems. 2015; pp 2224–2232.

(16) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* **2018**, *51*, 1281–1289.

(17) Corey, E.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166*, 178–192.

(18) Molyneux, A. J.; Cekirge, S.; Saatci, I.; Gál, G. Cerebral Aneurysm Multicenter European Onyx (CAMEO) trial: results of a prospective observational study in 20 European centers. *American Journal of Neuroradiology* **2004**, *25*, 39–51.

(19) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. *Organic Synthesis, Reactions and Mechanisms*; Springer, 1987; pp 19–73.

(20) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K., et al. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angewandte Chemie International Edition in English* **1993**, *32*, 201–227.

(21) Satoh, H.; Funatsu, K. 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**, *35*, 34–44.

(22) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.

(23) Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* **2014**,

(24) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science* **2018**, *9*, 6091–6098.

(25) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. A. Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation. *arXiv preprint arXiv:1811.02633* **2018**,

(26) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. Advances in neural information processing systems. 2017; pp 5998–6008.

(27) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **2017**, *3*, 1103–1113.

(28) Huczko, A. Template-based synthesis of nanomaterials. *Applied Physics A* **2000**, *70*, 365–376.

(29) Mann, M.; Nahar, F.; Schnorr, N.; Backofen, R.; Stadler, P. F.; Flamm, C. Atom mapping with constraint programming. *Algorithms for Molecular Biology* **2014**, *9*, 23.

(30) Fooshee, D.; Andronico, A.; Baldi, P. ReactionMap: An efficient atom-mapping algorithm for chemical reactions. *Journal of chemical information and modeling* **2013**, *53*, 2812–2819.

(31) Latendresse, M.; Malerich, J. P.; Travers, M.; Karp, P. D. Accurate atom-mapping computation for biochemical reactions. *Journal of chemical information and modeling* **2012**, *52*, 2970–2982.

(32) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications* **2019**, *10*, 1–11.

(33) Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. European Semantic Web Conference. 2018; pp 593–607.

(34) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**,

(35) Potdar, K.; Pardawala, T. S.; Pai, C. D. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications* **2017**, *175*, 7–9.

(36) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A comprehensive survey on graph neural networks. *arXiv preprint arXiv:1901.00596* **2019**,

(37) Luong, M.-T.; Pham, H.; Manning, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* **2015**,

(38) Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-First AAAI Conference on Artificial Intelligence. 2017.

(39) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* **2016**,

(40) Evgeniou, T.; Pontil, M. Regularized multi–task learning. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004; pp 109–117.

(41) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge, 2012.

(42) Edwards, W.; Barron, F. H. SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational behavior and human decision processes* **1994**, *60*, 306–325.

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

(44) Caruana, R. Multitask learning. *Machine learning* **1997**, *28*, 41–75.

(45) Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* **2017**,

(46) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Laino, T.; Reymond, J.-L. Data-Driven Chemical Reaction Classification, Fingerprinting and Clustering using Attention-Based Neural Networks. **2019**,

(47) Schneider, N.; Stiefl, N.; Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling* **2016**, *56*, 2336–2346.

(48) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579–2605.

(49) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D., et al. XSEDE: accelerating scientific discovery. *Computing in science & engineering* **2014**, *16*, 62–74.