Prediction of Chemical Reaction Type with Hierarchical Graph Neural Networks

Tatsuya Ishimoto[1]
ishimoto.t.ac@m.titech.ac.jp

Nobuyuki Yasuo[2]
yasuo.n.aa@m.titech.ac.jp

Masakazu Sekijima[1]
sekijima@c.titech.ac.jp

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[2] Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology, Tokyo, 152-8550, Japan

Keywords: Chemical reaction classification, Graph Neural Networks

Chemical reaction classification help chemists to index their reaction database and identify new types of reaction [1]. Data-driven reaction classification methods have been developed [2-3]. However, most of them do not adequately represent the hierarchical structure of chemical reactions. In this study, we propose the method of reaction classification with hierarchical graph representation. We experimented the model using the USTPO dataset to evaluate its efficiency.

[1] Bawden, D. Classification of Chemical Reactions: Potential, Possibilities and Continuing Relevance. *J. Chem. Inf. Comput. Sci.,* **1991**, *31 (2)*, 212–216.
[2] Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39–53.
[3] Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nature Machine Intelligence* **2021**, *3* (2), 144–152.

Molecular Optimization by Graph Generative Model using Transformers

Shunya Makino[1]
makino.s.ad@m.titech.ac.jp

Nobuaki Yasuo[2]
yasuo@cbi.titech.ac.jp

Masakazu Sekijima[1]
sekijima@c.titech.ac.jp

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[2] Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology, Tokyo, 152-8550, Japan

Keywords: machine learning, molecular optimization, graph generation, Transformers

In drug discovery, the significant cost and time required to develop new drugs have become a problem, and computational methods to reduce these costs have been actively researched. Lead optimization plays an important role in drug discovery and aims to enhance the activity or improve the ADMET properties of a compound known to be active against a target by changing its structure.

Recent advances in deep learning technology have led to the development of graph generation methods such as VAE and GAN[1, 2], which have been applied to the generation of molecular graphs. In the latest research, the Transformer[3] model has also been applied to graph representation learning, with excellent results in the task of predicting molecular properties.[4]

In this study, we extend graph representation learning with Transformer and propose a molecular graph generation method using the Transformer model and we test the model on the ZINC[5] database to demonstrate the usefulness of the proposed method.

[1] Simonovsky, Martin, and Nikos Komodakis. "Graphvae: Towards generation of small graphs using variational autoencoders." *International conference on artificial neural networks*. Springer, Cham, 2018.
[2] De Cao, Nicola, and Thomas Kipf. "MolGAN: An implicit generative model for small molecular graphs." *arXiv preprint arXiv:1805.11973* (2018).
[3] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
[4] Ying, Chengxuan, et al. "Do transformers really perform badly for graph representation?." *Advances in Neural Information Processing Systems* 34 (2021): 28877-28888.
[5] Irwin, John J., and Brian K. Shoichet. "ZINC– a free database of commercially available compounds for virtual screening." *Journal of chemical information and modeling* 45.1 (2005): 177-182.

An Enhanced Machine Learning Models for Predicting Retrosynthesis Accessibility

Mami Ozawa[1]
ozawa.m.ac@m.titech.ac.jp

Nobuaki Yasuo[2]
yasuo.n.aa@m.titech.ac.jp

Masakazu Sekijima[1]
sekijima@c.titech.ac.jp

[1] Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[2] Academy for Convergence of Materials and Informatics (TAC-MI), Tokyo Institute of Technology, Tokyo, 152-8550, Japan

Keywords: drug discovery, machine learning, synthesizability prediction, retrosynthesis, molecular generation

In the field of drug discovery, the amount of time and money spent on research and development is a problem. According to [1], it takes about as much as 10-15 years and 2.6 billion. In this context, the use of computational science to reduce the cost of drug discovery is attracting attention. An example of use is to generate new compounds using molecular generation models and filter them by predicting their synthetic accessibility using machine learning. RAscore[2] is one of the predictive models based on machine learning, and this study examines the practicality of the RAscore model and the problems associated with it. It also proposes a method for building models that can make more accurate predictions based on hypotheses about causes. Specifically, in addition to the ChEMBL[3] data used to train the original RAscore, the model was trained using data generated by a molecular generation model, ChemTS[4]. As a result, better area under the curve (AUC) and binary accuracy were achieved. The models, datasets, and source code are available at https://github.com/sekijima-lab/retrosynthesizability_prediction_models .

[1] Hansen RW DiMasi JA, Grabowski HG. *Innovation in the pharmaceutical industry: New estimates of r&d costs.* J Health Econ, 47(4):20–33, 2016.
[2] Amol Thakkar, Veronika Chadimov´a, Esben Jannik Bjerrum, Ola Engkvist, and Jean-Louis Reymond. *Retrosynthetic accessibility score (rascore) – rapid machine learned synthesizability classification from ai driven retrosynthetic planning.* Chem. Sci., 12:3339–3349, 2021.
[3] Nowotka M et al. Gaulton A, Hersey A. *The chembl database in 2017.* Nucleic Acids Res, 45(D1):D945–D954, 2017.
[4] Xiufeng Yang, Jinzhe Zhang, Kazuki Yoshizoe, Kei Terayama, and Koji Tsuda. *Chemts: an efficient python library for de novo molecular generation.* Science and Technology of Advanced Materials, 18(1):972–976, 2017.

Multi-Objective Molecular Optimization Using Monte Carlo Tree Search

Suzuki Takamasa[1 †]
suzuki.t.dq@m.titech.ac.jp

Ma Dian[1 †]
ma.d.ab@m.titech.ac.jp

Yasuo Nobuaki[2]
yasuo@cbi.titech.ac.jp

Sekijima Masakazu[1]
sekijima@c.titech.ac.jp

[1] Department of Computer Science, Tokyo Institute of Technology, 152-8550, Japan
[2] Academy for Convergence of Materials and Informatics, Tokyo Institute of Technology, Tokyo, 152-8550, Japan
[†] Corresponding Author

Keywords: In silico drug design, Molecular Generation, Multi-Objective Optimization

Drugs have saved millions of lives in human history, but the cost of developing a new drug has risen dramatically in the last decades. Recently, more and more computer-aided drug discovery (CADD) methods have been practiced in the pharmaceutical industry [1].

To address that problem, studies on molecular generation models which can generate compounds with novel structures have been raised in recent years. Early generative models decide their search direction on optimization for a single objective [3], however, a drug candidate required more multiple evaluations.

In this study, we developed a new deep learning-based extendable multiple-objective molecular generator, which could optimize molecules based on their properties and their affinity towards target proteins. This generator utilizes a recurrent neural network (RNN) to generate molecules and Pareto Multi-Objective Monte Carlo Tree Search (Pareto MOMCTS) to decide search direction.

By using our method, we have validated the generation of compounds for specific target proteins using the drug-like properties and docking scores as objective functions. In each search process, a new Monte Carlo search tree is built, and the change of the Pareto front, which decide the search direction, shows the transfer of the search process in different stages. This study will enable our method to provide more effective drug design assistance when compared to methods with a single objective, such as QED. However, compared with a single objective of drug-like properties, which made our method more effective in drug design.

[1] Stephen J. H.; Thomas U. M.; David T M.; Reza F.; Randall W. K.; Timothy J. M.; Stuart L S.; Dissecting cellular processes using small molecules: Identification of colchicine-like, taxol-like and other small molecules that perturb mitosis. *Chemistry & Biology,* **2000**, 7(4), 275–286.

[2] Rafael G. B.; Jennifer N. W.; David D.; José M. H. L.; Benjamín S. L.; Dennis S.; Jorge A. I.; Timothy D. H.; Ryan P. A.; Alan A. G.; Automatic chemical design using a data-driven continuous representation of molecules. ACS *Central Science*,**2018**, *4(2)*:268–276

[3] Robin W.; Floriane M.; Andreas S.; Hans B.; Frank N.; Djork-Arn´e C.; Efficient multi-objective molecular optimization in a continuous latent space. *Chemical Science*, **2019**, *10(34)*:8016–8024

Morphology-based drug effect profiling for delicate label-free phenotypic screening

Ryuji Kato[1]
kato-r@ps.nagoya-u.ac.jp

Kenjiro Tanaka[1]
tanaka-k@ps.nagoya-u.ac.jp

Yuto Takemoto[1]
takemoto.yuto.p2@s.mail.nagoya-u.ac.jp

Yuta Imai[1]
imai.yuta95@gmail.com

Yuto Okumura[1]
okumura.yuto@f.mbox.nagoya-u.ac.jp

Kei Kanie[1,2]
kanie-k@hiro.kindai.ac.jp

[1] Department of Basic Medicinal Sciences, Graduate School of Pharmaceutical Sciences, Nagoya University, Tokai National Higher Education and Research System, Furocho, Chikusa-ku, Nagoya, Aichi 464-8601, Japan
[2] Department of Biotechnology and Chemistry, Faculty of Engineering, Kindai University, 1 Umenobe, Takaya, Higashi-Hiroshima, Hiroshima 739-2116, Japan.

Keywords: Label-free, phenotypic screening, cell morphology,

By the advances of cell science, in vitro phenotypic screening has become a leading drug screening method. Label-free phenotypic analysis, which combines the recent image processing and machine learning technologies for quantitative cell image analysis, offers a balanced combination of high throughput and low cost, while providing a platform for non-invasive cell evaluation. It is not only cost-effective, but also flexible since label-free assay is free from complex validation and optimization of staining (such as customizing fluorescence wavelength overlap, bleaching effects, and transfection efficiency). Moreover, it is also free from the pre-knowledge of biomarkers.

In conventional label-free cell image analysis, there are basically two types of image data utilization concepts. The first type extracts multiple features from the cellular area in the image to describe cell morphology, such as area, roundness, and peripheral features [1], or intensity patterns and textures for the following analysis. Commonly, such features reflect the microscopically observed morphological characteristics, therefore analysis results, especially validity of the trained model, can be interpreted biologically. The second type uses image-tiles in the image as representative pixel patterns, and use them directly for training prediction models, such as deep learning models [2]. The collected image-tiles serve as functional descriptors, however, their features and how they were used in the model are not interpretable.

We have been proposing the importance of the former concept to establish explainable machine learning models for cell quality profiling. Practically, we have reported that morphology-based features from single cells can be summarized to describe the “population profile” of cells, and such features are important and effective for robust morphology-based cell analysis with various types of cells. In this presentation, we will present the performances and effectiveness of our morphology-based “cell population analysis” in the application of phenotypic drug screening [3].

[1] Sasaki, H. *et al.*: Label-free morphology-based prediction of multiple differentiation potentials of human mesenchymal stem cells for early evaluation of intact cells. *PLoS ONE,* **2014**, 9, e93952.
[2] Piotrowski, T. *et al.* Deep-learning-based multi-class segmentation for automated, non-invasive routine assessment of human pluripotent stem cell culture status. *Computers in Biology and Medicine,* **2021**, 129, 104172.
[3] Imai, Y. *et al.*: Label-free morphological sub-population cytometry for sensitive phenotypic screening of heterogenous neural disease model cells. *Scientific Reports*, **2022**, 12, 9296.

A recommendation algorithm for predicting drug effects considering directionality

Iori Azuma[1]
azuma-iori@g.ecc.u-tokyo.ac.jp

Tadahaya Mizuno[1, †]
tadahaya@mol.f.u-tokyo.ac.jp

Hiroyuki Kusuhara[1, †]
kusuhara@mol.f.u-tokyo.ac.jp

[1] Laboratory of Molecular Pharmacokinetics, Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, Japan
[†] Author to whom correspondence should be addressed

Keywords: Matrix factorization, Recommendation system, Side effects, Therapeutic indications

Considerable information is available regarding the safety, efficacy, and tolerance of approved drugs, which can minimize the expenditure and time required to predict new drug effects. With the development of machine learning techniques, recommendation systems based on matrix factorization for the prediction of drug effects have been well studied and various algorithms have been devised. However, most algorithms use only the presence or absence of known interactions as binary information and no studies that predict drug side effects using a recommendation system that considers the bilateral character of drug effects (i.e., a therapeutic effect and a side effect) are available. In the present study, we proposed a novel algorithm named neighborhood regularized bidirectional matrix factorization (NRBdMF) to predict drug effects by incorporating bidirectionality, which is a characteristic property of drug effects.

First, we conducted a survey of existing drug effect prediction algorithms using matrix factorization recommendation systems, focusing on their conceptual aspects. We found that existing methods are broadly classified into seven representative algorithms and employ binary information without considering the bilateral character of drug effects. Then we compared the prediction performance of these methods. Among the representative methods, neighborhood regularized logistic matrix factorization (NRLMF) [1] showed the best performance in benchmark tests.

Next, inspired by NRLMF, we developed a more generalized multilabel learning algorithm named NRBdMF that can handle bidirectionality and predict potential drug–target interaction. We used this proposed method for predicting side effects using a matrix that considered the bidirectionality of drug effects, in which known side effects were assigned a positive (+1) label and known treatment effects were assigned a negative (−1) label. We compared the side effect prediction performance between NRBdMF and NRLMF with enrichment score (ES), defined as the difference in enrichment values for each of the side effects and indications. The mean ES of NRBdMF (0.588 ± 0.101) outperformed that of NRLMF (0.191 ± 0.106). This indicates that the proposed method, NRBdMF, can enrich side effects and eliminate indications in the top-ranked predictions, which reduces the false positives and provides interpretable prediction results.

In a case study of the top 10 and bottom 10 of the predicted drug candidates that can cause hypertension, NRBdMF model could enrich more plausible drug candidates at the top ranks and less plausible candidates at the bottom ranks in predicting drug side effects.

These results indicate that the proposed NRBdMF achieves highly interpretable side effect prediction in the framework of recommendation systems based on matrix factorization.

[1] Y. Liu, M. Wu, C. Miao, P. Zhao, X.-L. Li, Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction, PLOS Comput. Biol. 12 (2016) e1004760