De novo generation of drug candidate compounds from disease-specific transcriptome data using deep learning

Chikashige Yamanaka[1]
yamanaka.chikashige215@mail.kyutech.jp

Shunya Uki[1]
uki.shunya677@mail.kyutech.jp

Kazuma Kaitoh[1]
kaito168@bio.kyutech.ac.jp

Yoshihiro Yamanishi[1]
yamani@bio.kyutech.ac.jp

[1] Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4, Kawazu, Iizuka, Fukuoka, 820-8502, Japan

Keywords: De novo drug design, gene expression profile, inverse correlation, variational autoencoder

It is extremely difficult to find molecules with desired properties in the drug discovery process. To tackle this problem, deep generative models, including variational autoencoder (VAE), have been receiving much attention. Most previous methods are based on chemical information, but they do not take into account biological information on genes and proteins. In this context, some previous studies have proposed deep generative models that generate active molecules for therapeutic target proteins using biological data such as gene expression profiles [1,2]; however, some limitations remain, such as difficulty in applying their models to diseases with unknown therapeutic targets. In a disease state, the disease-specific gene expression patterns typically emerge because of the disruption of homeostasis, and the disease can be treated by a drug molecule that invert the disease-specific gene expression patterns. Actually, the use of an inverse correlation between diseases and drugs is popular in drug repositioning [3,4], but the number of predictable candidate drugs is limited.

In this study, we propose to utilize the disease-specific gene expression profiles of patients toward the de novo design of drug candidate molecular structures. We present a computational model based on VAE with a transformer architecture to generate molecules that invert the disease-specific gene expression patterns. In the algorithm, we explore the latent space constructed by the VAE model using the inverse correlation with disease-specific transcriptome profiles of patients. We applied the model to several cancers and neurodegenerative diseases, and successfully generated new drug candidate molecular structures with the potential therapeutic effects against each disease. The proposed model can generate molecules with potential therapeutic effects against any disease in theory, thus the model is expected to make a great contribution to precision medicine.

[1] Méndez-Lucio, O.; Baillif, B.; Clevert, D.-A.; Rouquié, D.; Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence, *Nature Communications,* **2020**, 11, 10.

[2] Kaitoh, K.; Yamanishi, Y. TRIOMPHE: transcriptome-based inference and generation of molecules with desired phenotypes by machine learning, *Journal of Chemical Information and Modeling*, **2021**, 61, 4303–4320.

[3] Iorio, F.; Bosotti, R.; Scacheri, E.; Bernardo, D. D. Discovery of drug mode of action and drug repositioning from transcriptional responses, *Biophysics and computational biology*, **2010**, 107, 14621–14626.

[4] van Noort, V.; Schölch, S.; Iskar, M.; Zeller, G.; Ostertag, K.; Schweitzer, C.; Wenrner, K.; Weitz, J.; Koch, M.; Bork, P. Novel drug candidates for the treatment of metastatic colorectal cancer through global inverse gene-expression profiling, *Cancer Research*, **2014**, 74, 5690–5699.

O5-2

Data-Driven Understanding of TRPA1 Agonist Diversity

Kenjiro Tanaka[1]
tanaka.kenjiro.t8@s.mail.nagoya-u.ac.jp

Yuko Terada[2]
yukoterada@u-shizuoka-ken.ac.jp

Minami Matsuyama[2]
south59moco@gmail.com

Masaya Fujitani[1]
masaya.fujitani.0917@gmail.com

Masatoshi Shibuya[1]
m-shibu@ps.nagoya-u.ac.jp

Yoshihiko Yamamoto[1]
yamamoto-yoshi@ps.nagoya-u.ac.jp

Keisuke Ito[2]
sukeito@u-shizuoka-ken.ac.jp

Ryuji Kato[1]
kato-r@ps.nagoya-u.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 Food and Nutritional Sciences, Graduate School of Integrated Pharmaceutical and Nutritional Sciences, University of Shizuoka, 52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan

Keywords: Transient receptor potential ankyrin 1, Antagonist structure, Physicochemical property

Transient receptor potential ankyrin 1 (TRPA1) is a Ca^{2+}-permeable cation channel expressed in sensory neurons and non-neuronal cells of different tissues. TRPA1 has been linked to various physiological functions, including pain and pungency perception, energy metabolism, hormone secretion, and vasodilation [1,2], therefore it has been served as one of the amusing target of pharmaceutical targets. However, since TRPA1 function as a multimodal receptor, the structural diversity of TRPA1 agonists has not been fully understood.

We hypothesized that collecting a wider variation of TRPA1–compound interaction data would aid the understanding of its complex mechanism. Therefore, we aimed to explore such agonist diversity using an image-based TRPA1 assay system combined with an in silico chemical space clustering concept. We clustered our originally synthesized chemical library with 27 physicochemical molecular descriptors *in silico*, and selected structurally diverse compounds from each cluster for a detailed kinetic assay to investigate variations of agonist structural rules. Through two sets of assays evaluating various compounds in parallel with validating effects of the previously established structural rules, we discovered that different chemical groups contribute to agonist activity, indicating that there are multiple agonist design concepts. A novel core structure for a TRPA1 agonist has been also proposed.

Our data-driven approach to analyze the agonist diversity with physicochemical property rules facilitated the objective finding of the structural diversity and new rules in TRPA1 agonists [3].

[1] Mahajan, N.; Khare, P.; Kondepudi, K.K.; Bishnoi, M. TRPA1: Pharmacology, natural activators and role in obesity prevention. *European Journal of Pharmacology*, **2021**, *912*, 174553.
[2] Wang, Z.; Ye, D.; Ye, J.; Wang, M.; Liu, J.; Jiang, H.; Xu, Y.; Zhang, J.; Chen, J.; Wan, J. The TRPA1 channel in the cardiovascular system: Promising features and challenges. *Frontiers in Pharmacology*, **2019**, *10*, 1253.
[3] Terada, Y.; Tanaka, K.; Matsuyama, M.; Fujitani, M.; Shibuya, M.; Yamamoto, Y.; Ito, K.; Kato, R. Collection of data variation using a high-throughput image-based assay platform facilitates data-driven understanding of TRPA1 agonist diversity. *Applied Sciences*, **2022**, *12*, 1622.

O5-3

In Silico and In Vitro Screening for Inhibitors of SARS-CoV-2 Main Protease Avoiding Peptidyl Secondary Amides

Kazuki Yamamoto[1]
kazuki@ric.u-tokyo.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, Yokohama, 226-8501, Japan
[2] Academy for Convergence of Materials and Informatics, Tokyo Institute of Technology, Tokyo, 152-8550, Japan

Keywords: COVID-19, Protease inhibitor, Structure-based virtual screening

Vaccines and therapeutic agents are needed to control the damage caused by COVID-19 [1]. Currently available new drugs that inhibit SARS-CoV-2 main protease can suppress viral replication in the early stages of infection, but they are not easy to use in clinical practice because they involve inhibition of CYP3A4 [2-3]. Here, we performed structure-based virtual screening (SBVS) to find inhibitors with new scaffold structures, excluding compounds with peptidyl secondary amide bonds. The 180 compounds selected by SBVS were subjected to the main protease inhibition assay, and 9 compounds showed >5% inhibition at 20 μM. The IC50s of 6 of these compounds were determined by dose–response experiments and were on the order of 10^{-4} M. Based on these new scaffolds, we will try to optimize them.

[1] Robinson, Philip C., et al. COVID-19 therapeutics: Challenges and directions for the future. *Proceedings of the National Academy of Sciences* 119.15 (**2022**): e2119893119.
[2] Hammond, Jennifer, et al. "Oral nirmatrelvir for high-risk, nonhospitalized adults with Covid-19." *New England Journal of Medicine* 386.15 (**2022**): 1397-1408.
[3] https://www.mhlw.go.jp/content/11121000/000966645.pdf

O5-4

Active Learning via Incremental Revelation: Dipeptidyl Peptidase-4 Inhibitors Case Study

David Jimenez Barrero[1]
david.jimenez@elix-inc.com

Nazim Medzhidov[1]
nazim.medzhidov@elix-inc.com

[1]Elix Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: Active Learning, Drug Discovery, Machine Learning, Sitagliptin

Active Learning (AL) is a machine learning algorithm consisting of a constantly-improving model which is capable of achieving better accuracy with fewer labels than a regular model, as it is able to query the search-space and select expected promising samples[**1**]. This proves particularly useful in spaces which are too large and/or expensive to evaluate exhaustively, making it a clear proposition for use in the drug discovery context. Here the goal is to find a particular molecule with desired properties (high activity against a particular target, good synthesizability score, etc.) in a potentially infinite chemical space, and where the evaluation of each candidate is expensive economically as well as time consuming. Despite AL being theoretically a good fit for drug discovery, tests in this domain have been limited; due to the same economically-expensive nature of drug discovery. In this study, we designed a methodology that allows us to test performance of AL in a realistic framework by replicating real drug discovery campaigns while simulating the conditions human researchers face, and thus addressing common validity problems, such as having a limited and bounded search-space, and positively biased data. Under this scheme, we evaluated AL by replicating the discovery process of a dipeptidyl peptidase-4 (DPP4) inhibitor Sitagliptin [**2**]. We obtained comparable results to the estimated human performance by using both a single IC50 activity model and an ensemble of IC50 activity and lipophilicity models; thus proving the value of active learning in drug discovery, and the potential it has as a tool to aid medicinal chemists. We hope these landmark in-silico results pave the way for future tests of active learning using wet-lab experiments.

[1]Settles B. Active Learning Literature Survey. Computer Sciences Technical Report 1648. 2010 January 26; University of Wisconsin–Madison.

[2]Drucker, D., Easley, C. & Kirkpatrick, P. Sitagliptin. Nat Rev Drug Discov 6, 109–110 (2007).

SmilesFormer: Language Model for Molecular Design

Joshua Owoyemi[1]
joshua.owoyemi@elix-inc.com

Nazim Medzhidov[1]
nazim.medzhidov@elix-inc.com

[1] Elix Inc, Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081 Japan

Keywords: De novo drug design, Language model, Molecule Optimization

The objective of drug discovery is to find novel compounds with desirable chemical properties. Generative models have been utilized to sample molecules in the intersection of multiple property constraints. In this study we pose molecular design as a language modelling problem where the model implicitly learns the vocabulary and composition of valid molecules, hence it is able to generate new molecules of interest. We present a SmilesFormer, Transformer-based model [1] which is able to encode molecules, molecule fragments and fragment compositions as latent variables, which is in turn decoded to stochastically generate novel molecules. This is achieved by fragmenting the molecules into combinatorial groups, with methods such as RECAP [2], then learning the mapping between the input fragments and valid SMILES sequences. The model is able to optimize molecular properties through a stochastic latent space traversal technique [3]. This technique systematically searches the encoded latent space to find latent vectors that are able to produce molecules to meet the multi-property objective. The model was validated through various de novo molecular design tasks, achieving state-of-the-art performances when compared to previous methods. Furthermore, we used the proposed method to demonstrate a drug rediscovery pipeline for Donepezil [4], a known Acetylcholinesterase Inhibitor.

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. **2017**. 5998-6008.

[2] Xiao Qing Lewell, Duncan B. Judd, Stephen P. Watson, and Michael M. Hann. Recap - Retrosynthetic combinatorial analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *Journal of Chemical Information and Computer Sciences, 38(3)*. **1998**. 511–522.

[3] Jonas Mueller, David Gifford, and Tommi Jaakkola. Sequence to better sequence: continuous revision of combinatorial structures. *International Conference on Machine Learning*. **2017**. 2536–2544.

[4] Hachiro Sugimoto, Hiroo Ogura, Yasuo Arai, Youichi Iimura, and Yoshiharu Yamanishi. Research and development of donepezil hydrochloride, a new type of acetylcholinesterase inhibitor. *Japanese Journal of Pharmacology, 89(1)*. **2002**. 7–20.

O5-6

Hit to Lead Discovery of Benzylpiperidine Acetylcholinesterase Inhibitors Using Generative Models: a Retrospective Case Study

Nazim Medzhidov[1]
nazim.medzhidov@elix-inc.com

Joshua Owoyemi[1]
joshua.owoyemi@elix-inc.com

[1]Elix, Inc., Daini Togo Park Building 3F, 8-34 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan

Keywords: Generative Models, De novo drug design, Acetylcholinesterase inhibitors

Continuous increase in cost together with challenges associated with the traditional drug discovery process have facilitated the application of machine learning approaches in this domain. Past several years have provided with the examples where generative models proved successful in aiding de novo drug discovery [1]. However, the number of such success stories is still few and improved approaches are being continuously investigated. In this study, we used our in-house developed generative model [2] to cover a hit-to-lead discovery campaign of Acetylcholinesterase (AChE) inhibitors containing benzylpiperidine moiety. We designed a retrospective discovery scenario using publicly available data, where data was arranged in a chronological order and chemotype relevance. The goal was to replicate conditions of data scarcity during drug discovery process and investigation of novel target chemotype from an initial small set of hit compounds and public data. Generative model was then trained to generate molecules satisfying multi-objective criteria accounting for desired physical-chemical & medicinal chemistry properties, drug likeliness, synthetic accessibility, novelty, and activity towards AChE. Generated molecules were then subjected to an extensive post-processing pipeline and final list of suggested molecules was ranked and evaluated. Interestingly, among the final list of generated molecules we observed presence of documented successful compounds with benzylpiperidine moiety, having distant scaffolds to the training set. Moreover, an FDA-approved indanone-benzyliperidine AChE inhibitor donepezil and it's derivative [3] were also among the final list of suggested molecules generated. Our model was able to start from a set of hit compounds and successfully reach to novel and more potent scaffolds that were discovered at later stages of the original drug discovery campaign of AChE inhibitors.

[1] Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol* 37, 1038–1040 (**2019**).
[2] Owoyemi, J., *manuscript in preparation*
[3] Sugimoto, Hachiro et al. Research and development of donepezil hydrochloride, a new type of acetylcholinesterase inhibitor. *Japanese journal of pharmacology* 89,1 (**2002**)