Sanitize It Yourself: human-based sanitization checker against machine-generated chemical structures

Naruki Yoshikawa,[†,‡] Kentaro Rikimaru,[¶] and Kazuki Z Yamamoto[*,†,§]

†*SHaLX, Inc., Tokyo, Japan*

‡*Department of Computer Science, University of Toronto, Toronto, Canada*

¶*Preferred Networks, Inc., Tokyo, Japan*

§*Isotope Science Center, The University of Tokyo, Tokyo, Japan*

E-mail: kazuki@ric.u-tokyo.ac.jp

Phone: +81-3-4405-5628. Fax: +81-3-5841-2872

Abstract

Many computer-aided drug design (CADD) methods using deep learning have recently been proposed to explore the chemical space toward novel scaffolds efficiently. However, there is a tradeoff between the ease of generating novel structures and the chemical feasibility of structural formulas. To overcome the limitations of computational filtering, we have implemented a web application that allows easy compound sanitization by humans. The application is available at https://sanitizer.chemical.space/.

Introduction

Computer-aided drug design (CADD) has become an even more active research field with the rise of deep learning.[1] The cooperation of researchers from various backgrounds ranging from

organic chemistry to computer science is required to design feasible new compounds; however, it is not always easy to combine multidisciplinary insights. In recent years, there have been plenty of researches on molecular generative models. Still, some of these researches only look at numerical performance evaluations and lack the discussions about chemistry perspective of generated compounds.

In fact, a medicinal chemist Derek Lowe, known as a blogger influencer, posted a criticism article about the usefulness of molecular generative models. In the post, he pointed out even algorithm which claims to attain high-performance scores often generate chemically infeasible molecules, and he called such molecules as "crazy structures".[2] In order to find out such inappropriate structures, it is necessary to define and calculate the appropriateness of generated molecules. Some reports attempt at quantification of the appropriateness based on synthetic feasibility by automatic retrosynthesis tools.[3,4] If the tools successfully find reasonable synthetic routes for generated compounds, such compounds are considered to be synthetically feasible. These approaches can screen millions of generated compounds, but their reliability is controversial. Automatic retrosynthesis tools sometimes give incorrect synthetic routes for even simple molecules,[5] and there is room for further improvement. Given such situations, human-based sanitization of molecules is still necessary for ensuring the reliability of molecules generated by computers.

We have been advocating *social drug discovery*, in which we best-utilize the wisdom of the crowd for drug discovery. The power of the crowd on structure-based drug discovery was evaluated in our previous study.[6] In the study, Twitter users were asked to vote about which docking pose seems to be the most reasonable. The most voted pose matched the actual docking pose on 3 cases out of 3 questions. This result suggests that the majority opinion of chemists can be a useful source of information for drug discovery.

In this paper, we introduce the visualization tool for generated molecules on `chemical.space`[7] to enhance human-based sanitization of molecules. We first describe the problem of molecules generated by popular algorithms from the perspective of medicinal chemists, then introduce

the new visualization tools and finally describe the future perspective.

Problematic structures generated by molecular generation algorithms

The researches of generative models for molecules at the early days were mainly focused to increase the ratio of valid molecules among all generated ones. RDKit[8] has been used to assess the validity,[9] and validity is now regarded as one of the most important benchmarks for evaluating generative models .[10] After numerous efforts on increasing validity ratio, the generative models in most recent reports succeeded in achieving very high validity. However, it has been pointed out that such benchmark metrics including validity cannot properly evaluate the generated molecules. Renz and coworkers showed *failure mechanisms* of generative models. In the work, they exemplified the generated molecules could contain unstable, synthetically infeasible, or highly uncommon substructures.[11] In Figure 1, examples of such *unwanted molecules* are shown. In order to discard *unwanted molecules*, various quality filters have been proposed. For example, PAINS, and MCF filters are implemented in Moses packages.[12] Although these filters are useful to some extent, some unwanted molecules remain unfiltered because which substructures are *unwanted* depends on each user's individual situation. Therefore, users must prepare their own custom-defined filters to get meaningful generated molecules. In fact, REINVENT,[13] one of the most cited and widely used generative models, provides *Custom Alerts (CA)* component, which enables users to define their own *unwanted substructures*. Actual preparation of custom filters are laborious tasks, and tools for supporting visual inspection of users are essential to check and find out *unwanted molecules/substructures* among generated molecules.

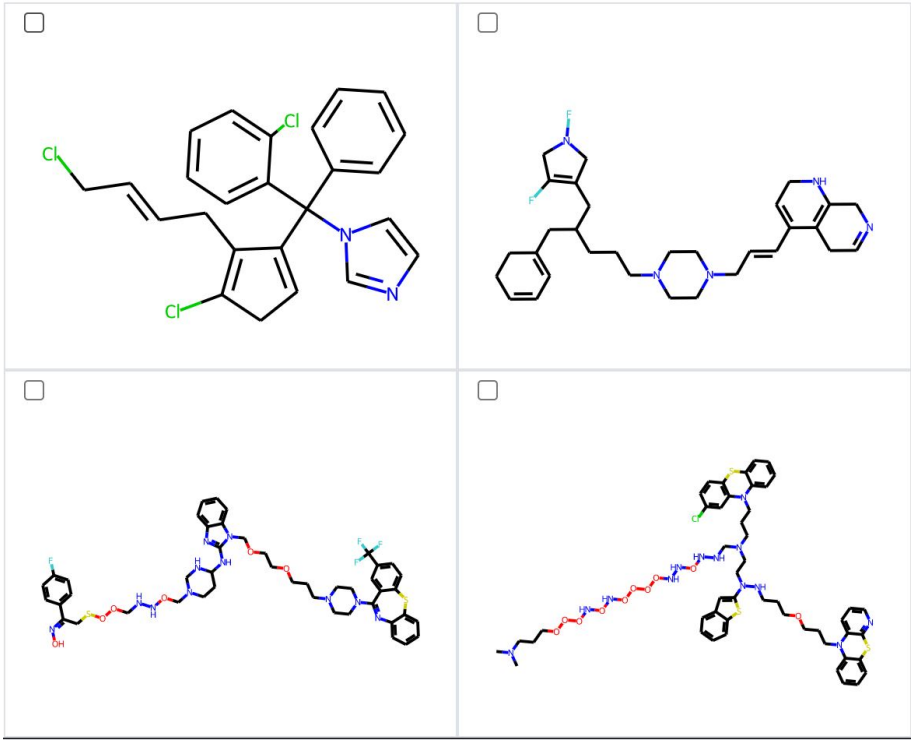

Figure 1: Examples of unwanted molecules. The molecules are taken from Ref[11] and depicted using our tool. Compounds at the top are created by a graph-based genetic algorithm. The left compound contains a reactive diene. The right one contains reactive nitrogen-fluorine, imine and diene moieties. Compounds at the bottom are generated by a SMILES-based LSTM. The two compounds contain long hetero-atom chains which are unstable and synthetically infeasible.

Molecule sanitization checker

We developed a web-based molecule visualization tool to enhance molecule sanitization through visual inspection. Organized visualization of chemical space of molecules is necessary to check the computer-generated molecules. Molecular scaffold is one of important characteristics where medicinal chemists pay attention to when checking molecules. Therefore, we adopted a substructure-based classification of molecules for visualization.

The workflow of the visualization is as follows: 1) The users upload a list of molecules in SMILES, 2) Subgraph mining[14] is conducted to find frequent substructures in the molecule list, 3) Frequent substructure list is shown to the user, and the user choose a substructure from the list, 4) Molecules containing the chosen substructure is shown to the user, and the user selects molecules of interest, 5) The user download the list of selected molecules. MoSS[15] was used as subgraph miner, RDKit[8] was used to visualize molecule. A screenshot of the visualizer is shown in Figure 2.

The web service is available at `https://sanitizer.chemical.space/`. If users want to deal with private data, they can build their own server using the source code for the service.

Case study

Evaluation of our web application was conducted by medicinal chemists. The goal of this case study was to find invalid molecules from molecules generated by one of the author's previous work[16] using this visualization tool. According to the users, unwanted molecules could easily be found from the list of generated molecules on this app. Example unwanted molecules are shown in Figure 3.

The users highlighted how easy it was to judge whether the selected substructure is useful or unwanted by looking at similar molecules with common structural moiety by the substructure-based classification function. We will continue to develop the application re-

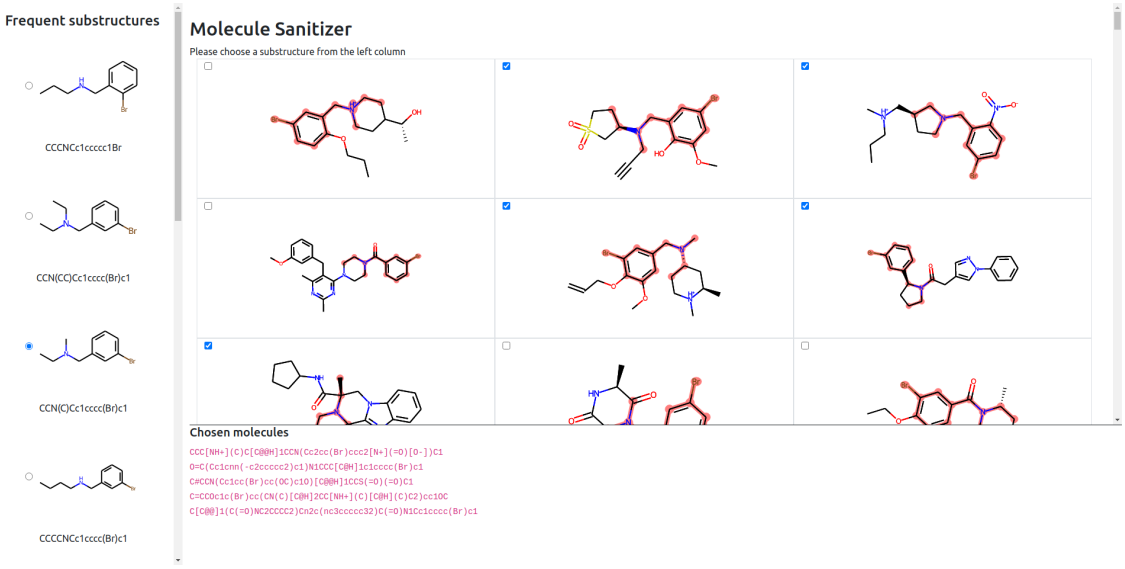

Figure 2: Molecule sanitization checker. The user selects a substructure from the left column, and check molecules that the user is interested in. The list of SMILES for checked molecules is shown in the lower column.

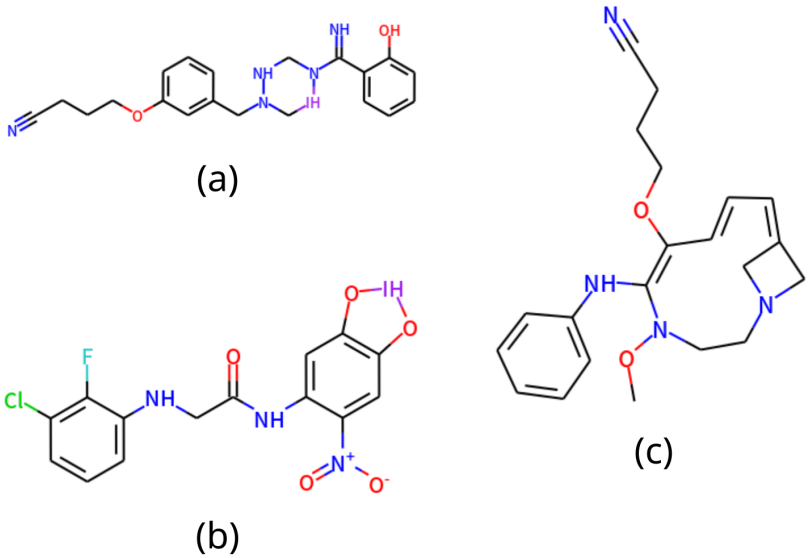

Figure 3: Found unwanted molecules generated by SMILES-based genetic algorithm.[16] (a) The compound contains nitrogen-iodine bonds, and the iodine atom has the inappropriate valency. (b) The compound contains oxygen-iodine bonds, and the iodine atom has the inappropriate valency. (c) The compound has unstable medium-sized ring.

flecting the users' opinion. Currently our tool now only supports automatically generated substructures. More functionality such as user-defined SMARTS filter would be useful to specify unwanted substructures more accurately.

Conclusion

In this study, we pointed out the problem of current generative models. It is very likely that some unwanted compounds are contained in generated molecules. Benchmark metrics are not sufficient to prioritize and select compounds in appropriate way from generated ones. It would be very useful if molecules with chemically unstable or synthetically infeasible substructures are captured effectively and automatically. Although there are attempts to filter out unwanted structures, visual inspection by experts is still necessary. We implemented a web-based visual inspection tool that takes advantages of common substructures found by subgraph mining. This will ease the validation of generated molecules from wide-range of people. We will expand this tool for crowd-based drug discovery in the future by implementing voting functions.

Acknowledgement

The authors thank Ryuichi Kubo for technical support, and Ryuichiro Ishitani, and Masaaki Kotera for helpful discussions.

Data and Software Availability

The molecule visualizer is available at `https://sanitizer.chemical.space/`, and the source code is available in `https://github.com/n-yoshikawa/molecule-sanitizer` under MIT license.

Supporting Information Available

The SMILES list used in the case study was extracted from `https://github.com/tsudalab/ChemGE/blob/master/results/log-rdock-chemlet`. The result is available at `https://sanitizer.chemical.space/viewer/demo`.

References

(1) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

(2) Generating Crazy Structures. `https://blogs.sciencemag.org/pipeline/archives/2020/09/30/generating-crazy-structures`, Accessed 29 March 2021.

(3) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *Journal of Chemical Information and Modeling* **2020**, *60*, 5714–5723, PMID: 32250616.

(4) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021**, *12*, 3339–3349.

(5) Borrelli, W.; Schrier, J. Evaluating the Performance of a Transformer-based Organic Reaction Prediction Model. *ChemRxiv* **2021**,

(6) Yamamoto, K. Social Drug Discovery Project (in Japanese). *Digital Practice* **2018**, *9*, 842–858.

(7) Yoshikawa, N.; Kubo, R.; Yamamoto, K. Z. Twitter integration of chemistry software tools. *Journal of Cheminformatics* **2021**, *13*, 46.

(8) RDKit: Open-source cheminformatics. `http://www.rdkit.org`.

(9) Rafael, G.-B.; 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 Cent. Sci.* **2018**, *4*, 268–276.

(10) Brown, N.; Marco, F.; H.S., S. M.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *Journal of Chemical Information and Modeling* **2019**, *59*, 1096–1108.

(11) Renz, P.; Rompaey, D. V.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies* **2019**, *32-33*, 55–63.

(12) Polykovskiy, D. et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Frontiers in Pharmacology* **2020**, *11*, 565644.

(13) Blaschke, T.; Arús-Pous, J.; Chen, H.; Margreitter, C.; Tyrchan, C.; Engkvist, O.; Papadopoulos, K.; Patronov, A. REINVENT 2.0: An AI Tool for de Novo Drug Design. *Journal of Chemical Information and Modeling* **2020**, *60*, 5918–5922.

(14) Mrzic, A.; Meysman, P.; Bittremieux, W.; Moris, P.; Cule, B.; Goethals, B.; Laukens, K. Grasping frequent subgraph mining for bioinformatics applications. *BioData Mining* **2018**, *11*, 20.

(15) Borgelt, C.; Meinl, T.; Berthold, M. MoSS: A Program for Molecular Substructure Mining. Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations. New York, NY, USA, 2005; p 6–15.

(16) Yoshikawa, N.; Terayama, K.; Sumita, M.; Homma, T.; Oono, K.; Tsuda, K. Population-based de novo molecule generation, using grammatical evolution. *Chemistry Letters* **2018**, *47*, 1431–1434.

Graphical TOC Entry