SOFTWARE

Sanitize It Yourself: Web-based molecular sanitization for machine-generated chemical structures

Naruki Yoshikawa[1], Kentaro Rikimaru[2] and Kazuki Z Yamamoto[3,4,5*]

*Correspondence:
kazuki@ric.u-tokyo.ac.jp
[3]Isotope Science Center, The University of Tokyo, Tokyo, Japan
Full list of author information is available at the end of the article

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-based software in which users can share and evaluate computer-generated compounds. The web service is available at https://sanitizer.chemical.space/.

Keywords: Molecular generative model; Chemical space; Molecular sanitization; Visualization; Molecule editor; Knowledge sharing; Open science

ORCIDs

- Naruki Yoshikawa: 0000-0003-1546-8709, `@narukiyoshikawa`
- Kentaro Rikimaru: 0000-0001-5106-5523
- Kazuki Z Yamamoto: 0000-0001-6231-0475

1 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 chemical perspectives of generated compounds. Extracting candidates of true value for drug development from computationally generated molecules requires a multifaceted evaluation. [2, 3].

In fact, the usefulness of molecular generative models has been criticized by some medicinal chemists, in which they point out even algorithms claiming to attain high-performance scores often generate chemically infeasible molecules, and such molecules are referred to as "crazy structures." [4] In order to find out such inappropriate structures, it is necessary to define and calculate the appropriateness of generated molecules. Some researchers attempt to quantify the appropriateness based on synthetic feasibility by running automatic retrosynthesis tools and counting the number of synthetic steps required to produce the molecule [5, 6]. These approaches can screen millions of generated compounds, but their reliability is not

established yet, since automatic retrosynthesis tools sometimes give incorrect synthetic routes for even simple molecules [7]. Given such situations, human-based molecular sanitization is still necessary for ensuring the reliability of molecules generated by computers.

Decision-making in drug development often relies on the tacit knowledge of experienced medicinal chemists in the field [8], which renders it useful to gather a wide range of their views [9]. In this context, we have been proposing using collective knowledge through social web as one of the trends in open science for drug discovery, which we call *social drug discovery*. In tasks such as prioritizing compounds [10] or selecting the feasible 3D interaction structures in structure-based drug design process, we can harness the power of the crowd through means such as voting [11]. The majority opinion of scientists can be a valuable source of information for drug discovery. Likewise, we can apply the wisdom of crowds [12] to scrutinize computer-generated molecules.

In this paper, we introduce a web-based molecular sanitization tool for computer generated molecules. Users can share a list of generated molecules in the service and ask for an evaluation of their generative algorithm from a wide range of people. This web-based voting function democratizes the drug discovery research process and may facilitate social drug discovery. In the following chapters, 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.

2 Problematic structures generated by molecular generation algorithms

The researches of generative models for molecules in the early days were mainly focused to increase the ratio of valid molecules among all generated ones. RDKit [13] has been used to assess validity [14], and validity is now regarded as one of the most important benchmarks for evaluating generative models [15]. After numerous efforts on increasing the validity ratio, the generative models in the 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 [16]. In Figure 1, examples of such *unwanted molecules* are shown.

In order to discard *unwanted molecules*, various quality filters have been proposed. For example, Pan Assay Interference Compounds [17] (PAINS), and medicinal chemistry filters (MCF) are implemented in MOSES packages [18]. 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 [19], 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 is a laborious task, and tools for supporting visual inspection of users are

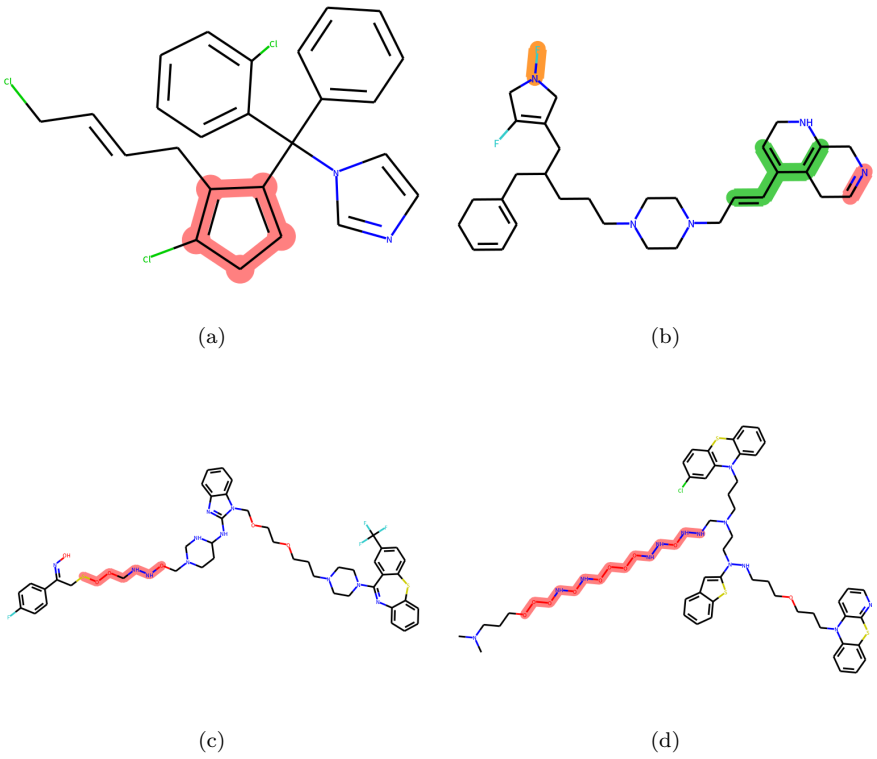


Figure 1 Examples of unwanted molecules. The molecules are taken from Ref [16] and depicted using our tool. Compounds (a) and (b) are created by a graph-based genetic algorithm. The compound (a) contains a reactive diene, and (b) contains reactive nitrogen-fluorine (orange), imine (red) and diene moieties (green). Compounds (c) and (d) are generated by a SMILES-based LSTM. The two compounds contain long hetero-atom chains which are unstable and synthetically infeasible.

essential to check and find out *unwanted molecules/substructures* among generated molecules.

3 Web-based molecular sanitization

We developed a web-based molecule visualization tool to enhance molecular sanitization through visual inspection. Organized visualization of the chemical space of molecules is necessary to check the computer-generated molecules. In addition, we encourage users to share their molecule lists with the world and vote for promising molecules to gain the wisdom of crowds [20] or collective human intelligence [21]. The web service is available at `https://sanitizer.chemical.space/`. Although the main goal of this project is to share molecules for collective knowledge, users can build their own server using the source code for the service if they want to deal with private data.

3.1 Implementation

Our website provides a graphical voting system for posted molecules. To implement the voting system, we need to store information of (1) molecules posted by users, (2) supplementary information on molecules, and (3) molecule evaluation provided

by users. We used Django [22] for the backend, and jQuery [23] and RDKit for JavaScript [24] for frontend visualization.

User management and login We ask users to log in to the service to identify evaluators. We currently support OAuth authorization with Twitter or ORCID using Python Social Auth [25]. If the user is not logged in, they are treated as a guest user.

Project Users can create a new project or view existing projects on the dashboard page (Figure 2a). A project is a unit to control a list of molecules. The users can create a new project by uploading up to 10000 molecules in SMILES (Figure 2b) in one project. The uploaded molecules are processed by RDKit [13] to add tags on the server side. We currently support three tags: Rule of five [26], PAINS [17], and MCF [18]. Rule of five (Ro5) tag is added when the molecular property satisfies all of the following criteria: (i) the molecule has no more than 5 hydrogen bond donors, (ii) the molecule has no more than 10 hydrogen bond acceptors, (iii) the molecular mass is less than 500 daltons, and (iv) the calculated log P is less than 5. PAINS filter is implemented using RDKit's `FilterCatalog`, and MCF is implemented using the SMARTS list from the MOSES benchmark [18].

Molecule Viewer On the project page, the users can see the list of uploaded molecules. The molecule is rendered by RDKit for JavaScript on the client side. The users can evaluate molecules by pushing `like` and `dislike` buttons. They can filter molecules using tags (Ro5, PAINS, and MCF) or current users' evaluations (like or dislike). If the link to the project page is shared, multiple users can evaluate the molecules. They can export the SMILES of filtered or evaluated molecules.

Substructure Search Users can search molecules which contain specific substructures using SMARTS (SMILES arbitrary target specification) [27]. They can edit SMARTS with JSME [28] or input in the text box (Figure 3a). Molecule editor can be invoked from the menu button of each molecule to search for similar molecules. The searched substructure is highlighted on the viewer page.

Nearest Neighbor Search Users can also search for similar molecules a specific molecule inside a project. Nearest neighbors are determined based on angular distance between MACCS (Molecular ACCess System) keys [29]. An approximate nearest neighbors search library Annoy [30] is used for fast neighbor search.

Molecule Info Each molecule has a separate page to check the detailed information. The page contains the information of the molecular property to determine whether it satisfies the rule of five, the name of users who evaluated the molecule, and the information of which substructure is filtered by PAINS (if applicable). This page has two buttons for Twitter integration: to share the molecule in Twitter, and to send it to the retrosynthesis bot [31] (Figure 3c).

3.2 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 authors' previous work [32] using this visualization tool. According to the users, unwanted molecules could easily be found from the list of generated molecules thanks to the visualization and searching functionalities on this app. An example of unwanted molecules is shown in Figure 4.

4 Conclusion

In this study, we pointed out the problem of current molecular 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 an appropriate way from generated ones. It would be very helpful 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 tool that eases molecular sanitization based on the visual inspection of experts. We will continue to develop the application reflecting the users' opinions.

Availability and requirements

- Project name: Sanitize It Yourself
- Project repository: https://github.com/n-yoshikawa/molecule-sanitizer
- Operating system(s): Platform independent
- Programming languages: Python and JavaScript
- Other requirements: RDKit, Django
- License: MIT License

Availability of data and materials
The website is available at https://sanitizer.chemical.space/.

Competing interests
The authors declare that they have no competing interests.

Funding
The server maintenance cost has been and will be paid by SHaLX Inc.

Authors' contributions
NY implemented the software. NY, KR, and KZY wrote the paper. KZY maintains the cloud web servers.

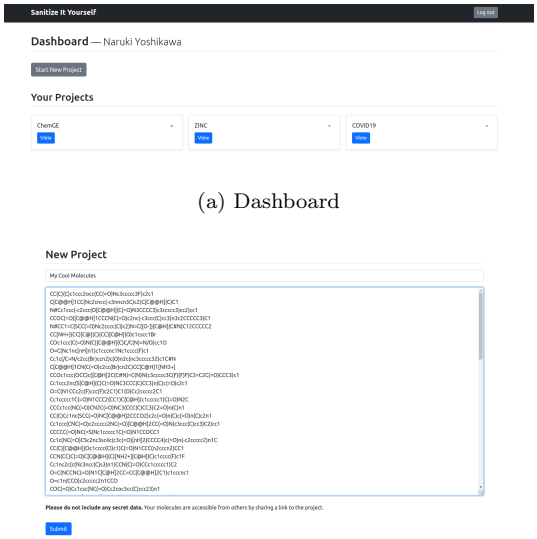
Acknowledgements
The authors thank Ryuichi Kubo for technical support, and Ryuichiro Ishitani, and Masaaki Kotera for helpful discussions.

Author details
[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2], Preferred Networks, Inc., Tokyo, Japan. [3]Isotope Science Center, The University of Tokyo, Tokyo, Japan. [4], SHaLX Inc., Tokyo, Japan. [5]Department of Computer Science, Tokyo Institute of Technology, Yokohama, Japan.

References

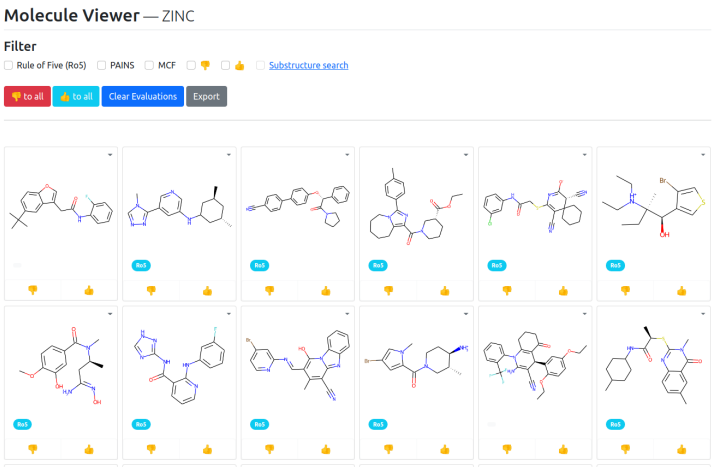
1. Sanchez-Lengeling, B., Aspuru-Guzik, A.: Inverse molecular design using machine learning: Generative models for matter engineering. Science **361**(6400), 360–365 (2018). doi:10.1126/science.aat2663
2. Walters, W.P., Murcko, M.: Assessing the impact of generative ai on medicinal chemistry. Nature Biotechnology **38**(2), 143–145 (2020)
3. Walters, W.P., Barzilay, R.: Critical assessment of ai in drug discovery. Expert Opinion on Drug Discovery, 1–11 (2021)
4. Generating Crazy Structures. https://blogs.sciencemag.org/pipeline/archives/2020/09/30/generating-crazy-structures. Accessed 29 March 2021.

5. Gao, W., Coley, C.W.: The synthesizability of molecules proposed by generative models. Journal of Chemical Information and Modeling **60**(12), 5714–5723 (2020). doi:10.1021/acs.jcim.0c00174
6. 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. **12**, 3339–3349 (2021). doi:10.1039/D0SC05401A
7. Borrelli, W., Schrier, J.: Evaluating the performance of a transformer-based organic reaction prediction model. ChemRxiv (2021). doi:10.33774/chemrxiv-2021-3nqv9
8. Gomez, L.: Decision making in medicinal chemistry: The power of our intuition. ACS Med. Chem. Lett. **9**(10), 956–958 (2018)
9. Mayweg, A., Hofer, U., Schnider, P., Agnetti, F., Galley, G., Mattei, P., Lucas, M., Boehm, H.-J.: ROCK: the roche medicinal chemistry knowledge application – design, use and impact. Drug Discov. Today **16**(15), 691–696 (2011)
10. Kutchukian, P.S., Vasilyeva, N.Y., Xu, J., Lindvall, M.K., Dillon, M.P., Glick, M., Coley, J.D., Brooijmans, N.: Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. PLoS One **7**(11), 48476 (2012)
11. Yamamoto, K.: Social drug discovery project (in Japanese). Digital Practice **9**(4), 842–858 (2018). http://id.nii.ac.jp/1001/00191483/
12. Hack, M.D., Rassokhin, D.N., Buyck, C., Seierstad, M., Skalkin, A., ten Holte, P., Jones, T.K., Mirzadegan, T., Agrafiotis, D.K.: Library enhancement through the wisdom of crowds. Journal of Chemical Information and Modeling **51**(12), 3275–3286 (2011). doi:10.1021/ci200446y
13. RDKit: Open-source cheminformatics. http://www.rdkit.org
14. 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. **4**(2), 268–276 (2018). doi:10.1021/acscentsci.7b00572
15. 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 **59**(3), 1096–1108 (2019). doi:10.1021/acs.jcim.8b00839
16. Renz, P., Rompaey, D.V., Wegner, J.K., Hochreiter, S., Klambauer, G.: On failure modes in molecule generation and optimization. Drug Discovery Today: Technologies **32-33**, 55–63 (2019). doi:10.1016/j.ddtec.2020.09.003
17. Baell, J.B., Holloway, G.A.: New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. Journal of Medicinal Chemistry **53**(7), 2719–2740 (2010). doi:10.1021/jm901137j
18. Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A., Zhavoronkov, A.: Molecular sets (MOSES): A benchmarking platform for molecular generation models. Frontiers in Pharmacology **11** (2020). doi:10.3389/fphar.2020.565644
19. 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 **60**, 5918–5922 (2020). doi:10.1021/acs.jcim.0c00915
20. Hack, M.D., Rassokhin, D.N., Buyck, C., Seierstad, M., Skalkin, A., ten Holte, P., Jones, T.K., Mirzadegan, T., Agrafiotis, D.K.: Library enhancement through the wisdom of crowds. Journal of Chemical Information and Modeling **51**(12), 3275–3286 (2011). doi:10.1021/ci200446y
21. Cincilla, G., Masoni, S., Blobel, J.: Individual and collective human intelligence in drug design: evaluating the search strategy. Journal of Cheminformatics **13**(1), 80 (2021). doi:10.1186/s13321-021-00556-6
22. Django Software Foundation: Django. https://djangoproject.com
23. The jQuery Team: jQuery. https://jquery.com/
24. RDKit for JavaScript (Official). https://github.com/rdkit/rdkit/tree/master/Code/MinimalLib
25. Matías Aguirre: Python Social Auth. https://python-social-auth.readthedocs.io/
26. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J.: Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews **23**(1-3), 3–25 (1997). doi:10.1016/s0169-409x(00)00129-0
27. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html. Accessed 2021-11-22
28. Bienfait, B., Ertl, P.: JSME: a free molecule editor in JavaScript. Journal of Cheminformatics **5**(1), 24 (2013). doi:10.1186/1758-2946-5-24
29. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of mdl keys for use in drug discovery. Journal of Chemical Information and Computer Sciences **42**(6), 1273–1280 (2002). doi:10.1021/ci010132r
30. Annoy: Approximate Nearest Neighbors in C++/Python. https://github.com/spotify/annoy. Accessed: 2021-11-21
31. Yoshikawa, N., Kubo, R., Yamamoto, K.Z.: Twitter integration of chemistry software tools. Journal of Cheminformatics **13**(1), 46 (2021). doi:10.1186/s13321-021-00527-x
32. Yoshikawa, N., Terayama, K., Sumita, M., Homma, T., Oono, K., Tsuda, K.: Population-based de novo molecule generation, using grammatical evolution. Chemistry Letters **47**(11), 1431–1434 (2018). doi:10.1246/cl.180665



(a) Dashboard

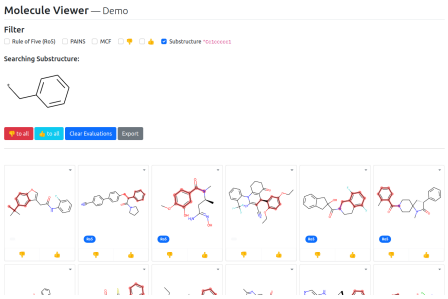
(b) New Project



(c) Molecule Viewer

Figure 2 Screenshots of the website. (a) The dashboard page is shown after the user is logged in. They can create a new project or see previous project. (b) A list of SMILES is provided by the user to create a new project. Each SMILES are separated by new lines. (c) Users can evaluate molecules by pushing like and dislike buttons.

(a) SMARTS editor



(b) Substructure search

(c) Molecule info

Figure 3 Screenshots of the website (continued). (a) Users can search molecules with substructure. (b) The detailed information of molecules can be checked in the molecule info page.

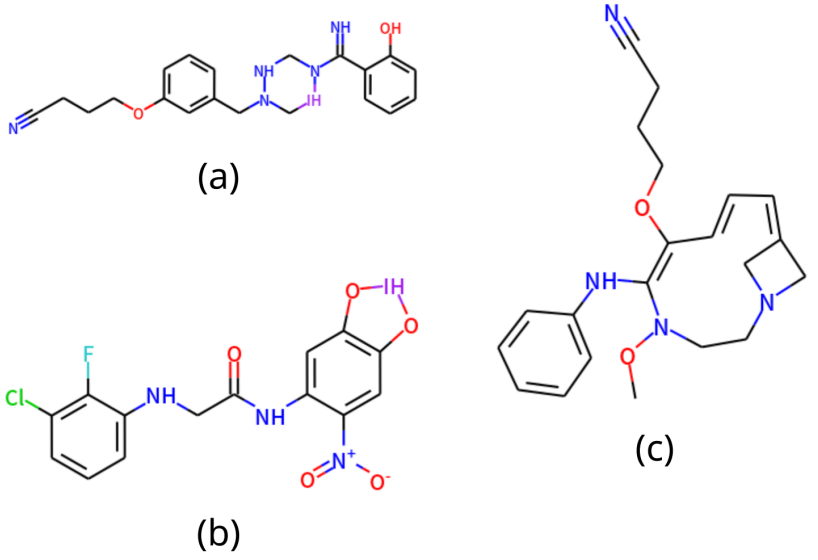


Figure 4 Identified unwanted molecules generated by SMILES-based genetic algorithm [32]. (a) The compound contains nitrogen-iodine bonds, and the iodine atom has the inappropriate valency. (b) The compound contains oxygen-iodine bonds, and the iodine atom has the inappropriate valency. (c) The compound has unstable medium-sized ring.