Machine learning for computationally designing synthesizable molecule

2021年6月18日　統計数理研究所　オープンハウス

Qi Zhang　総合研究大学院大学　統計科学専攻　博士課程5年

1, Introduction

Computationally molecular design is a computer-aided way of designing molecules. Of which the object is to generate novel molecular compounds with desirable properties.

1.1, Categories of computationally molecular design

- Virtual screening

Step1, Generate a large virtual molecule library;
Step2, Select candidates from the library based on demand.

- De novo molecular design

Step1, Generate a set of molecules;
Step2, Score the generated molecules;
Step3, Search for better molecules with respect to the score.

1.2, Recent progress in de novo molecular design

- Gómez-Bombarelli, R et al. maps the molecules into a continuous space by variational autoencoder, score the molecules by a neural networks regression model and perform a gradient based optimization in the continuous space[1].
- Segler, M et al. models the distribution of existing molecules by recurrent neural networks, optimize the distribution towards the target properties by finetuning with the promising molecules[2].
- Jin, W et al. use junction tree to encode the molecular structure, embedded vectors with high predicted scores will be decoded to the molecular structures[3].

1.3, Consideration of synthesizability

The majority of previous work do not explicitly account for synthetic feasibility, without which will lead to molecules that are very difficult to synthesize or potentially unstable. To solve this problem, synthesis is incorporated into the optimization steps. by constructing virtual molecules with building blocks(reactants) via synthesis prediction, synthesis instructions will be kept, thus the synthesizability.

- Brad-shaw et al. proposed a model for reactant selection, and synthesis the product by molecular transformer[4].
- Guo et al. proposed a Bayesian algorithm for designing synthetic path given a target molecule[5],
- Korovina et al. proposed a Bayesian optimization strategy for molecule selection based on target properties, molecules are synthesized by reaction prediction[6].

1.4, Our goal

We propose a de novo design method, which aims at solving the following remaining problems in this field.

- *Limitations of the number of synthetic steps*. Virtual synthesis is a resource-intensive process which causes the most time cost of the overall design workflow, more serious as increasing the complexity of synthetic route. Such that most works only focus on few steps synthesis, which limit the searching space.
- *Synthetic route should be designed artificially.* The structures of synthetic path is very unsettled, previous attempts of ducking this issue by fix the synthetic route with a relatively simple structure, only optimize the reactants among it. However, it is not reasonable assuming both chemical reaction involves the same number of reactants and the same process of synthesis. So far, there is no "non manual intervention" way to generate the synthetic path.
- *Searching space is subject to the existing building blocks.* Previous work design molecules by selecting reactants from the existing reactant pool, the searching space is highly bounded by the existing data, others break through the boundary by modifying the substructures of the existing molecules, however, there is no suggestion on how to synthesis the newly modified molecules.

2, Methods

2.1, Outline

Suppose a collection of commercially available molecules is accessible, denoted as $\boldsymbol{S} = \{s_i\}_{i=1}^{N}$ where s_i is a single molecule. A trained virtual synthetic model defines the mapping $s = F_{VS}(S)$, where $S \subset \boldsymbol{S}$. Note that F_{VS} is an unfixed composite function convolves a set of single-step virtual synthetic models f_{VS} define the mapping $s = f_{VS}(s_1, s_2), s_i \in S$. A trained score model defines the mapping $y = f_{score}(s)$, this mapping shall be vary based on demand, however, a larger y indicates a more desirable s in this work. Following these definitions, the ultimate goal of de novo design is to find out all possible S, such that $(f_{score} \circ F_{VS})(S) \geq y^*$, where, $(a \circ b)(x)$ means "evaluate b at x, then evaluate a at the result $b(x)$", y^* is a pre-designed target score.

2.2, Asynchronous de novo design

One of the concepts behind our proposal is the separation of design and synthesis procedure. In general, molecular has to be synthesized for scoring. To keep the design procedure going without obtaining the product, instead of optimizing each candidates with respect to the real score, we employ a surrogate model, which maps the reactant set directly to a surrogate score. Since there's no need to wait for product prediction, the computational time of design procedure dramatically decreased. This also makes the virtual synthesis procedure able to be processed parallelly.

2.3, Pool enrichment

Another concepts behind out proposal is to continuously add newly synthesized molecules into the reactant pool, called pool enrichment. Adding new building blocks into the pool expends the searching space, which brings more chance for designing promising molecules.

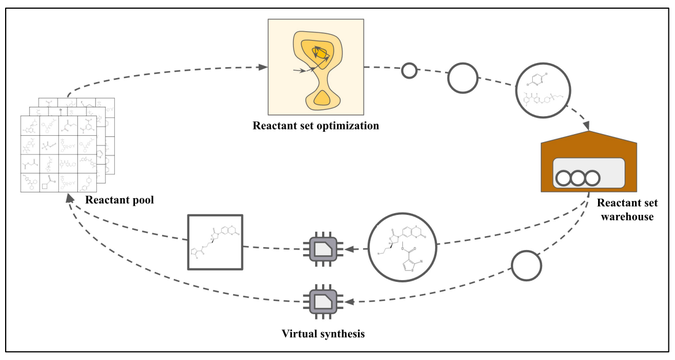


Illustration of proposed method.

3, Experiment

To demonstrate the usefulness of this work, we compared the performance with a conventional de novo design pipeline which incorporates a synthesis step followed by scoring in each iteration. The tasks are as follow:

Target properties:

- Dielectric constant < 3.8
- Glass transition temperature >450

Initial reactant pool:

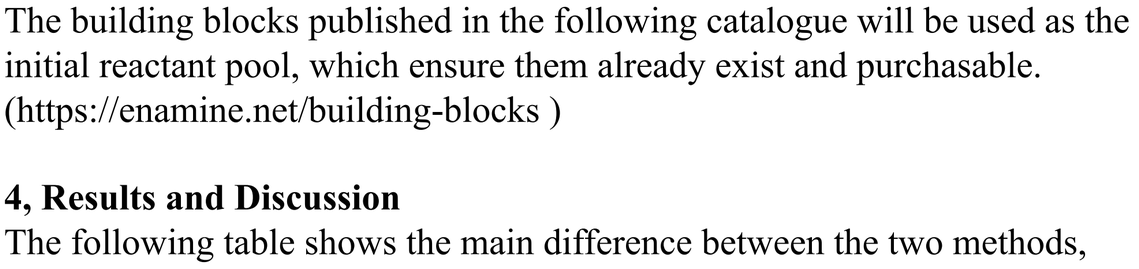
The building blocks published in the following catalogue will be used as the initial reactant pool, which ensure them already exist and purchasable. (https://enamine.net/building-blocks)

4, Results and Discussion

The following table shows the main difference between the two methods, benefited from the pool enrichment procedure, our method is able to explore wider searching space, which leads to 5 times more promising candidates. By implementing optimization and synthesis separately, our method decreased more than 40% of computational time.

Method	Conventional de novo	Proposed method
Number of promising candidates per step	25(±16)	137(±37)
Time cost per step	41.7(±2.2)sec.	24(±1.4)

At the same time, without any pre-designed synthetic path structures, our method is able to generate synthetic path with different structures.



Designed molecules with synthetic path.

Reference

[1] Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276.
[2] Segler, Marwin HS, et al. "Generating focused molecule libraries for drug discovery with recurrent neural networks." ACS central science 4.1 (2018): 120-131.
[3] Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." International Conference on Machine Learning. PMLR, 2018.
[4] Bradshaw, John, et al. "A model to search for synthesizable molecules." arXiv preprint arXiv:1906.05221 (2019).
[5] Guo, Zhongliang, et al. "Bayesian Algorithm for Retrosynthesis." Journal of Chemical Information and Modeling 60.10 (2020): 4474-4486.
[6] Korovina, Ksenia, et al. "Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.